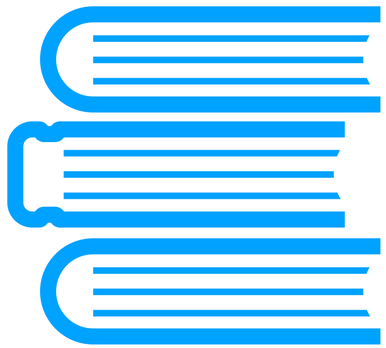
Génération de texte à partir de connaissances

Claire Gardent
CNRS/LORIA

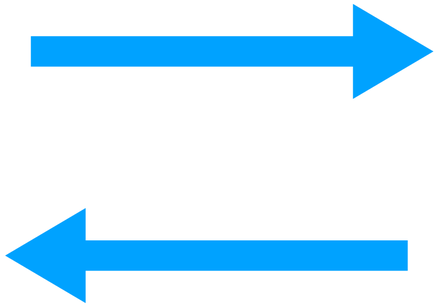
Chaire IA xNLG “*Generating from Multiple Sources into Multiple Languages*”

Traitement automatique des langues



Analyse

(Lire)



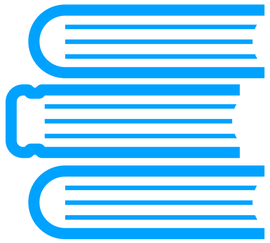
Génération

(Écrire)

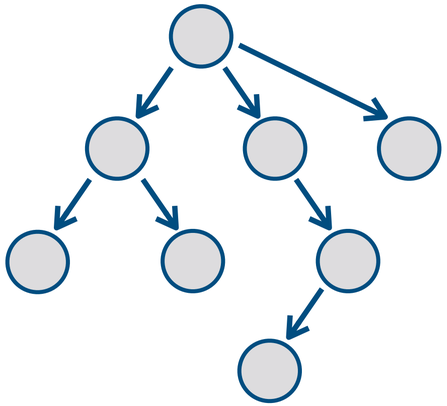
Génération de texte

Les entrées

Bases de données, données tabulaires



Textes



Graphe

Génération de texte

Et aussi à partir de …

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A group of young people playing a game of frisbee.

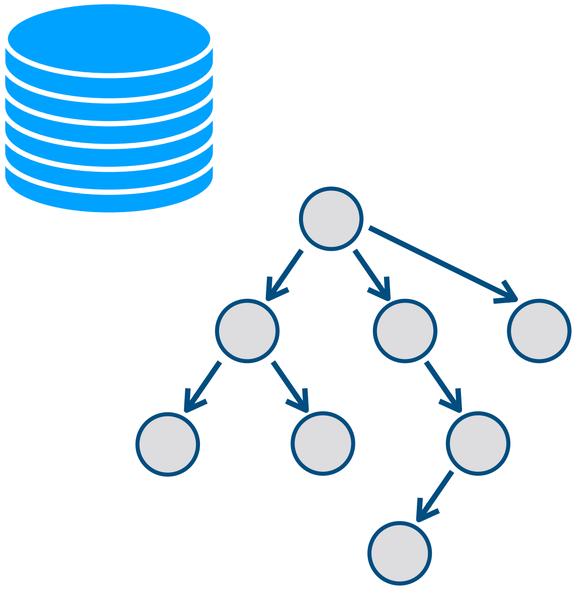
Two hockey players are fighting over the puck.

Images

Vidéos

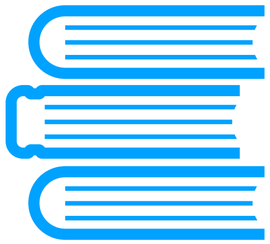
Génération de texte

Les objectifs : Pourquoi faire ?



Verbaliser

- Un graphe
- Une BD

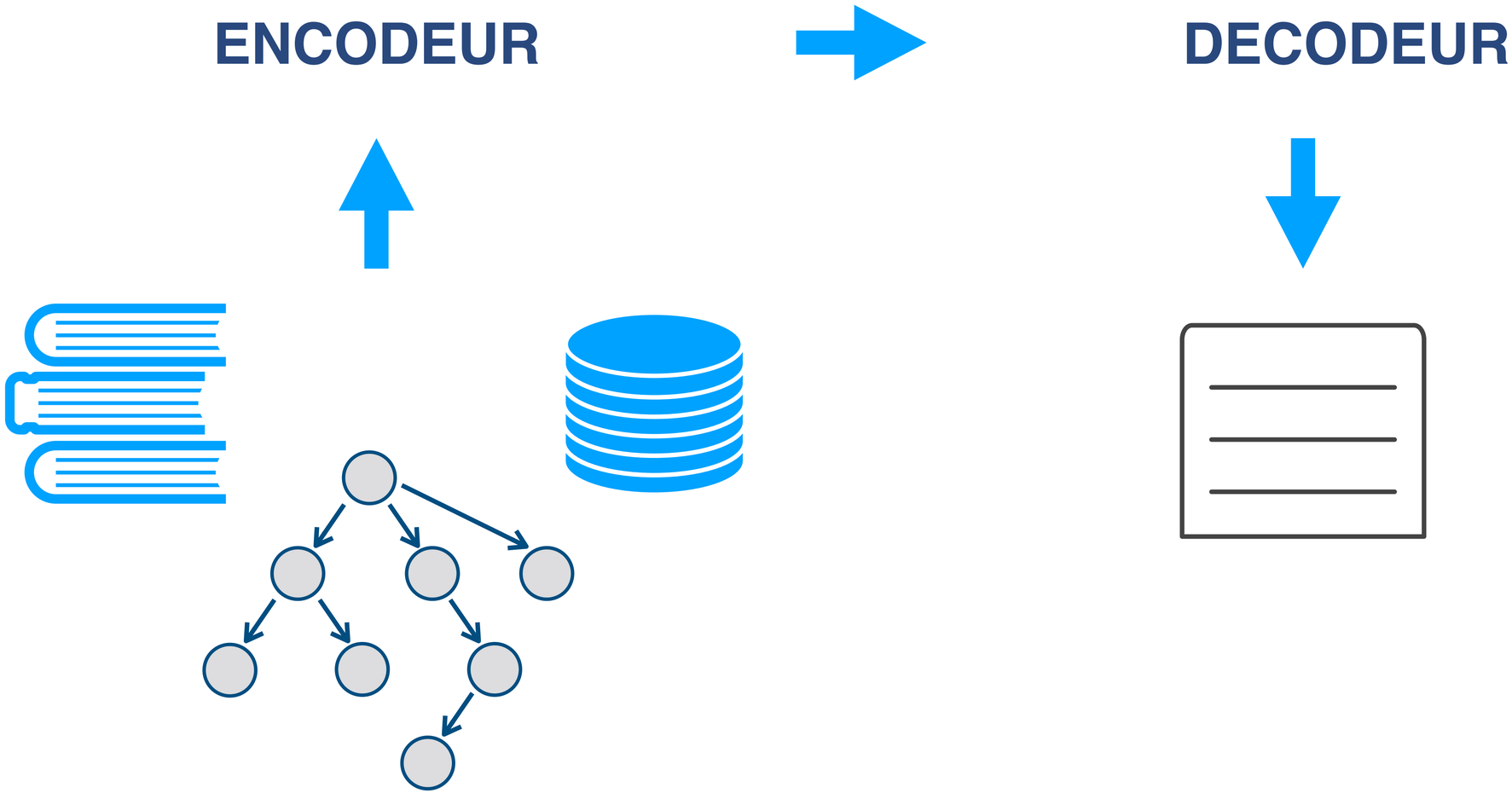


Simplifier, résumer, paraphraser

- Un/des texte(s)

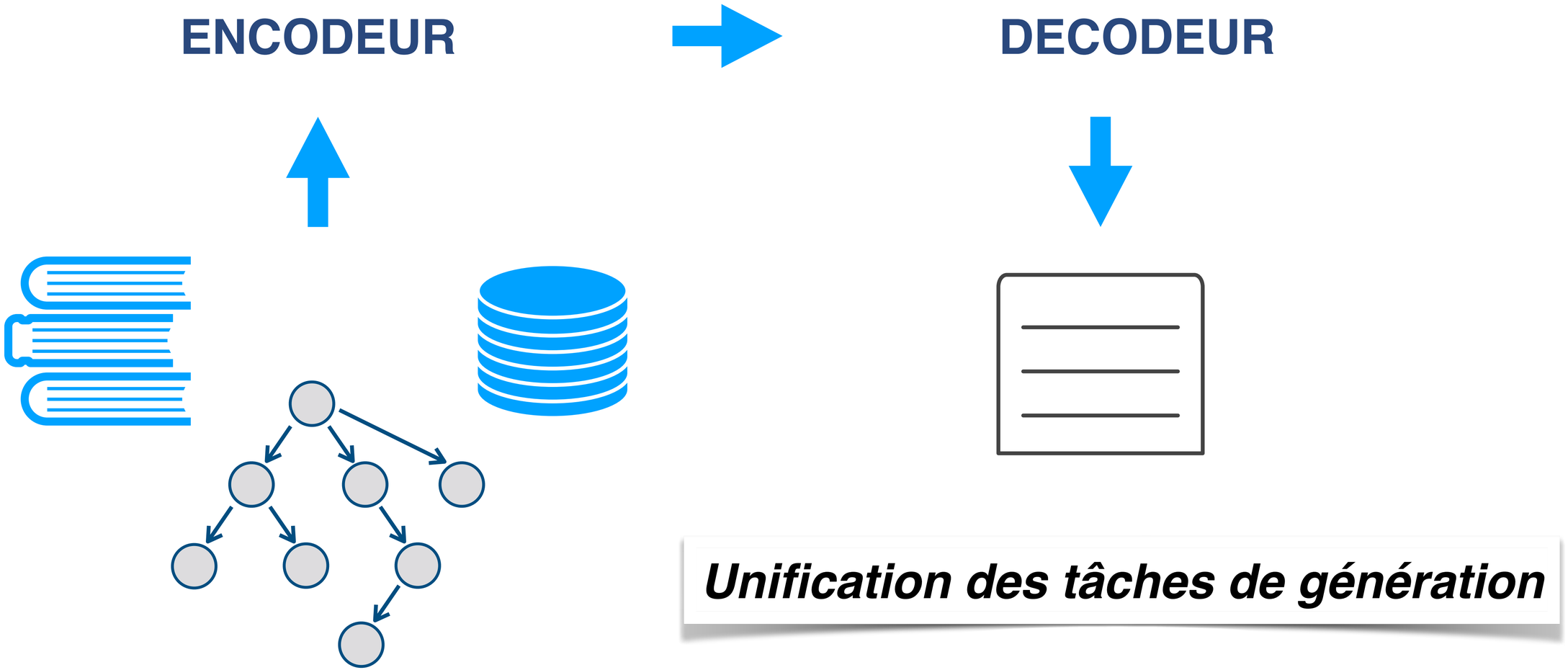
Génération de texte conditionnelle

Entrée -> Texte



Génération de texte conditionnelle

Entrée -> Texte



Unification des tâches de génération

Génération conditionnelle de texte

A partir

- d’une représentation sémantique

*Abstract Meaning Représentions (AMR) -> **21 langues** UE*

- d’un graphe de connaissances

*Resource Description Format (RDF) -> **Langues peu dotées***
(Breton, Galois, Irlandais)

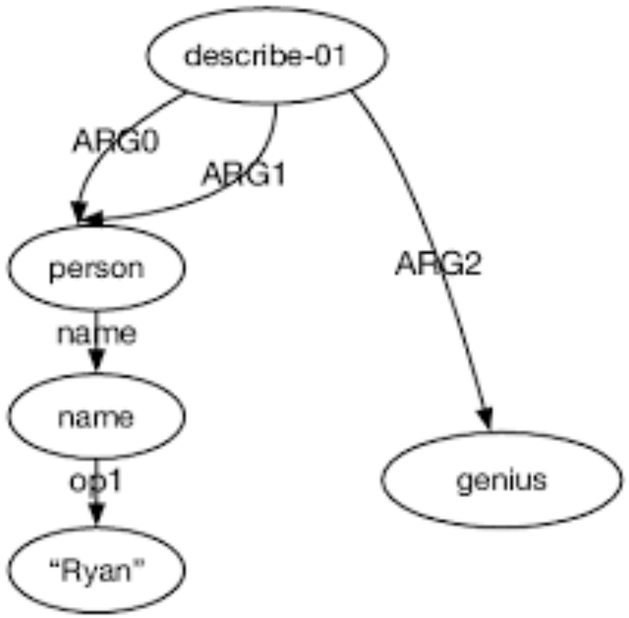
- de textes

Génération de biographies Wikipedia
Biais de données

Génération multilingue à partir de graphes de représentations sémantiques abstraites

Angela Fan and Claire Gardent
“Multilingual AMR-to-Text Generation”
EMNLP 2020.

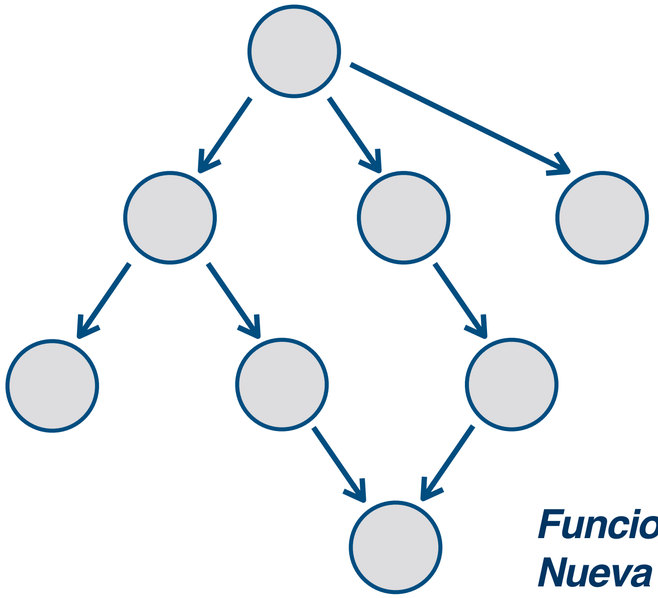
Graphe AMR (Abstract Meaning Representation)



- Graphe acyclique
- Noeuds : concepts
- Arcs : roles sémantiques

Ryan describes himself as a genius

Graphe AMR —> 21 Langues



Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

US officials held an expert group meeting in January 2002 in New York.

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Langues romanes, germaniques, slaves, ouraliennes

Enjeux

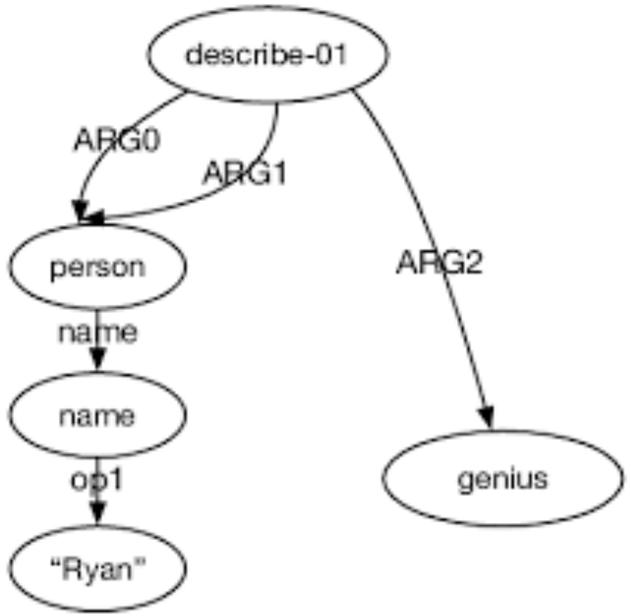
- Graphe —> Sequence

Enjeux

- Graphe —> Sequence
- Peu de données d’apprentissage

Enjeux

- Graphe —> Sequence
- Peu de données d’apprentissage
- Graphe sous-spécifié

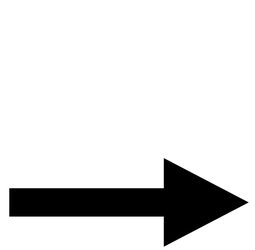


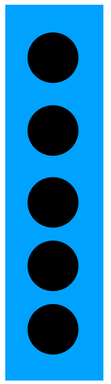
Enjeux

- Graphe —> Sequence
- Peu de données d’apprentissage
- Graphe sous-spécifié
- Génération multilingue : les langues cibles ont une syntaxe et une morphologie variées

Encodeur

Graphe

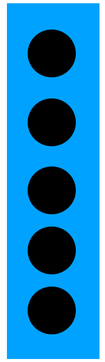




Représentation vectorielle

Encodeur

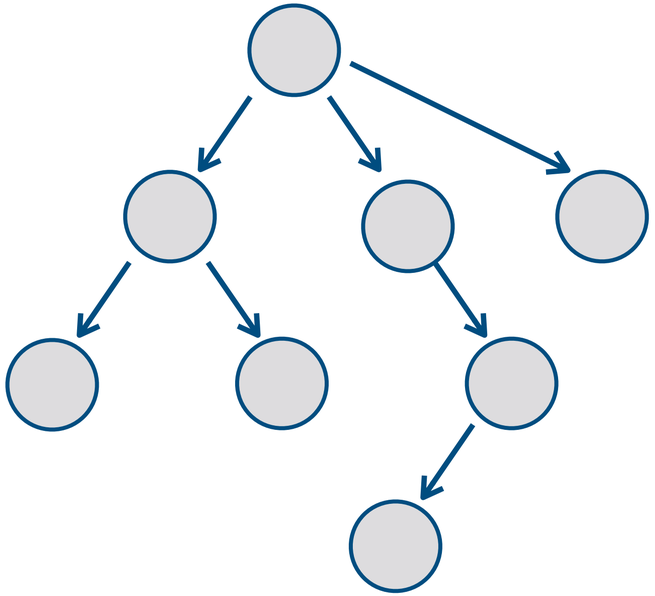
Graphe



Représentation vectorielle

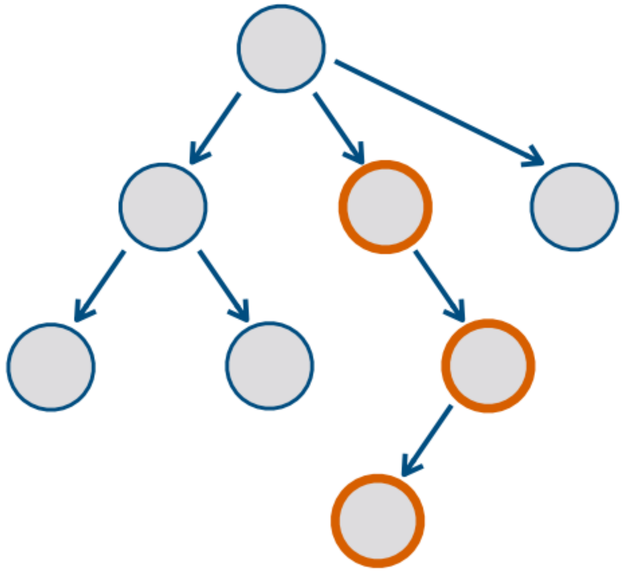
- Linéarisation
- Plongements structurels
- Sous mots
- Pré-apprentissage (MLM)

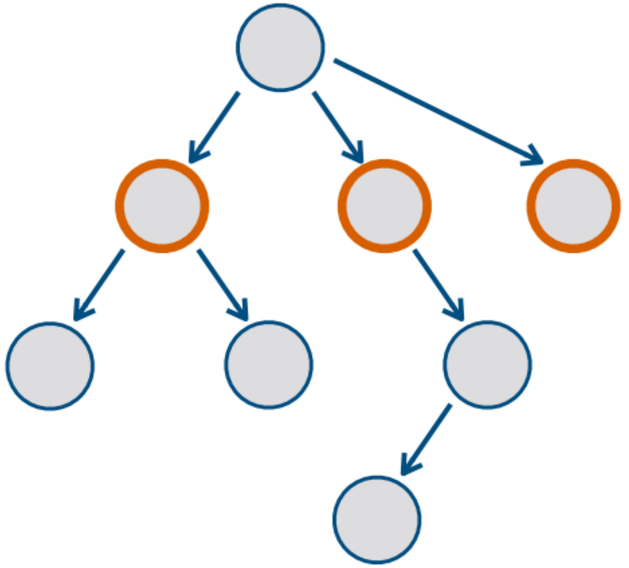
Linéarisation



hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2
States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

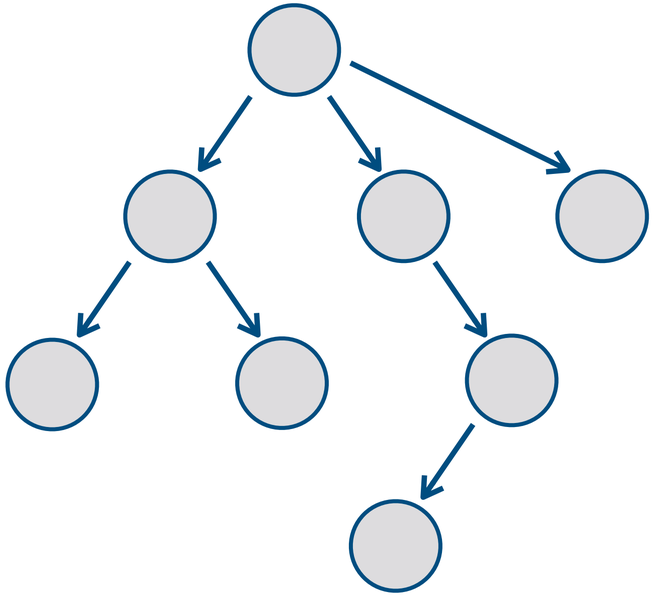
Ajout de plongements positionnels pour les noeuds soeurs et les branches





Pré-apprentissage

- Sur des AMRs “argent”
- 30M phrases tirées de CCNET et analysées avec JAMR



Décodeur
(Génération)

Texte

Représentation vectorielle

Génération multilingue

- Plongements cross-lingues XLM
- Modèle de langue pré-entrainé (30M phrases) pour chaque langue cible
- Encodeur-décodeur multilingue

French

Des responsables américains

….

Spanish

Funcionarios estadounidenses

….

Slovak

Americkí predstavitelia

….

Bulgarian

Американските служители

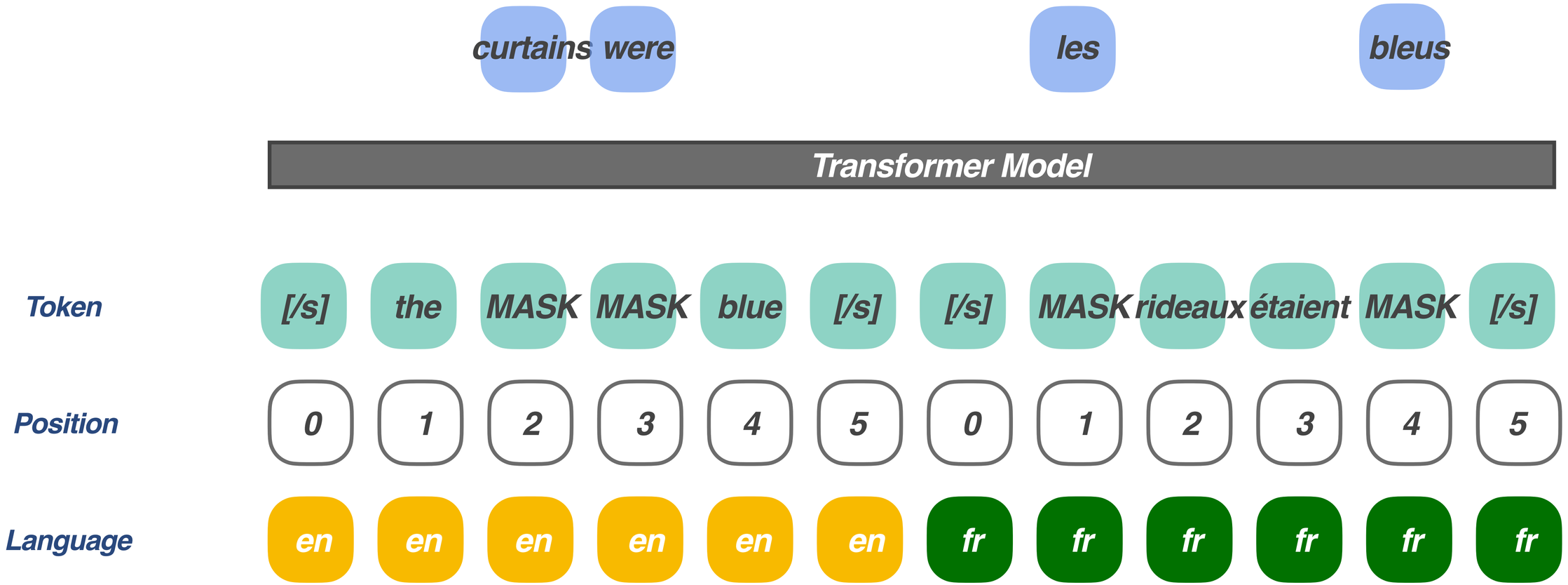
….

Swedish

Amerikanska tjänstemän

…

Plongements cross-lingues



Cross-lingual Language Model Pretraining

Guillaume Lample, Alexis Conneau

Encodeur-décodeur multilingue

Génération en slovaque

sv

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Americký predstavitelia usporiadali stretnutie expertnej skupiny v 2002 v New Yorku.

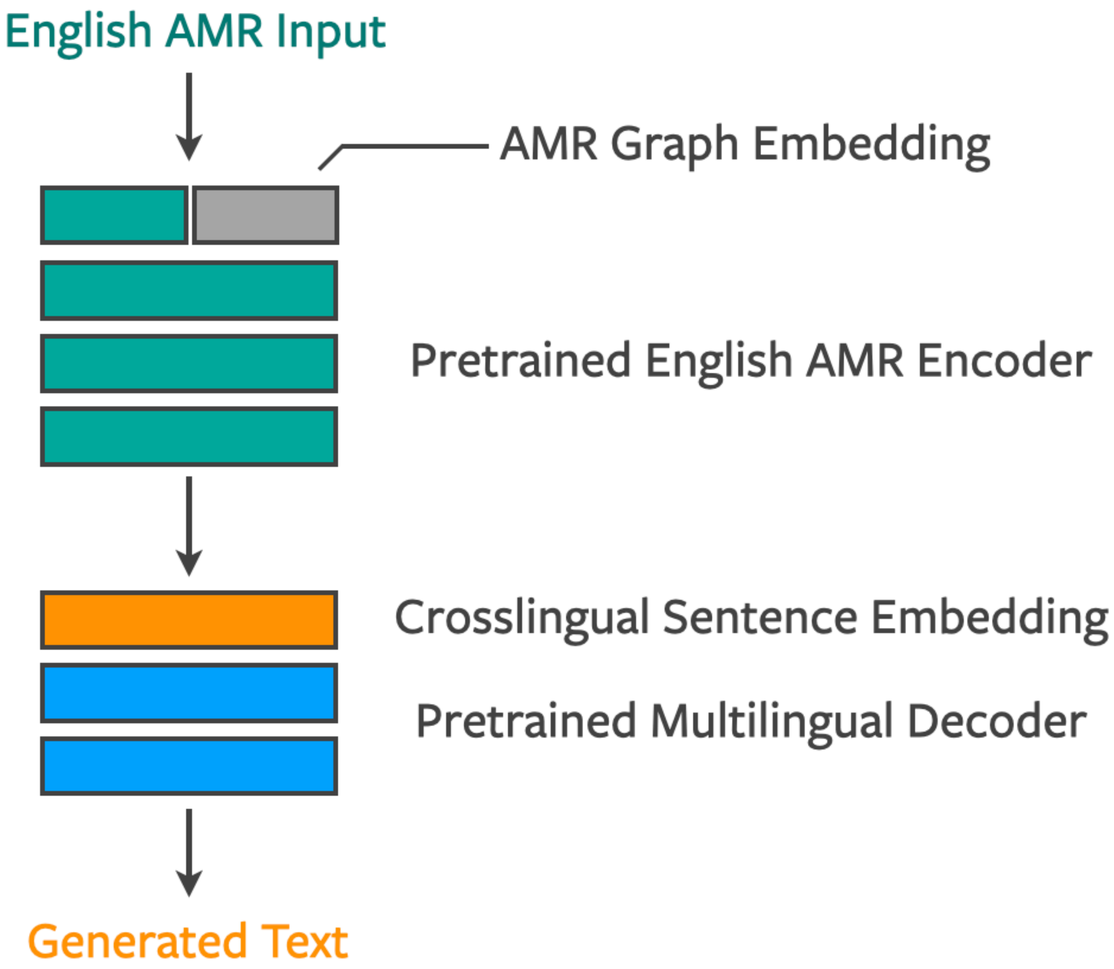
Génération en Français

fr

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Modèle AMR-Texte multilingue



- Encodeur: pre-apprentissage d'AMRs
- Décodeur: pre-apprentissage de LMs, modèle multilingue

Données d'apprentissage (21 langues)

hold

:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official

:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group

:time date-entity :year 2002 :month 1

:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.	Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.	Americkí predstavitelia usporiadali stretnutie expertnej skupiny v 2002 v New Yorku.	Американските служители проведоха среща на експертна група през януари 2002 г. в Ню Йорк.	Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.
French	***Spanish***	***Slovak***	***Bulgarian***	***Swedish***

Données

- Apprentissage
  - AMR argent, Texte Europarl, 21 langues

- Evaluation
  - AMR argent, Texte Europarl, 21 langues
  - AMR or, Texte LDC. 3 langues
  - Espagnol, allemand, italien

Evaluation

Automatique (BLEU)

- Ablation
- Comparaison avec deux modèles de base (baseline)
- Impact des langues d’apprentissages

Human-Based

- Word-Order, Morphology, Semantic adequacy, Paraphrasing

Ablation

Modèle de base (Anglais, BLEU)	32.5
+ Plongements de graphes	32.9
+ Plongements crosslingues	33.0
+ Encodeur pré-entrainé	33.4
+ Decodeur pré-entrainé	33.8

https://github.com/jflanigan/jamr

Comparaison: Monolingue v. Multilingue

Monolingue

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1
United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-
of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparaison: Monolingue vs. Multilingue

Monolingue

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

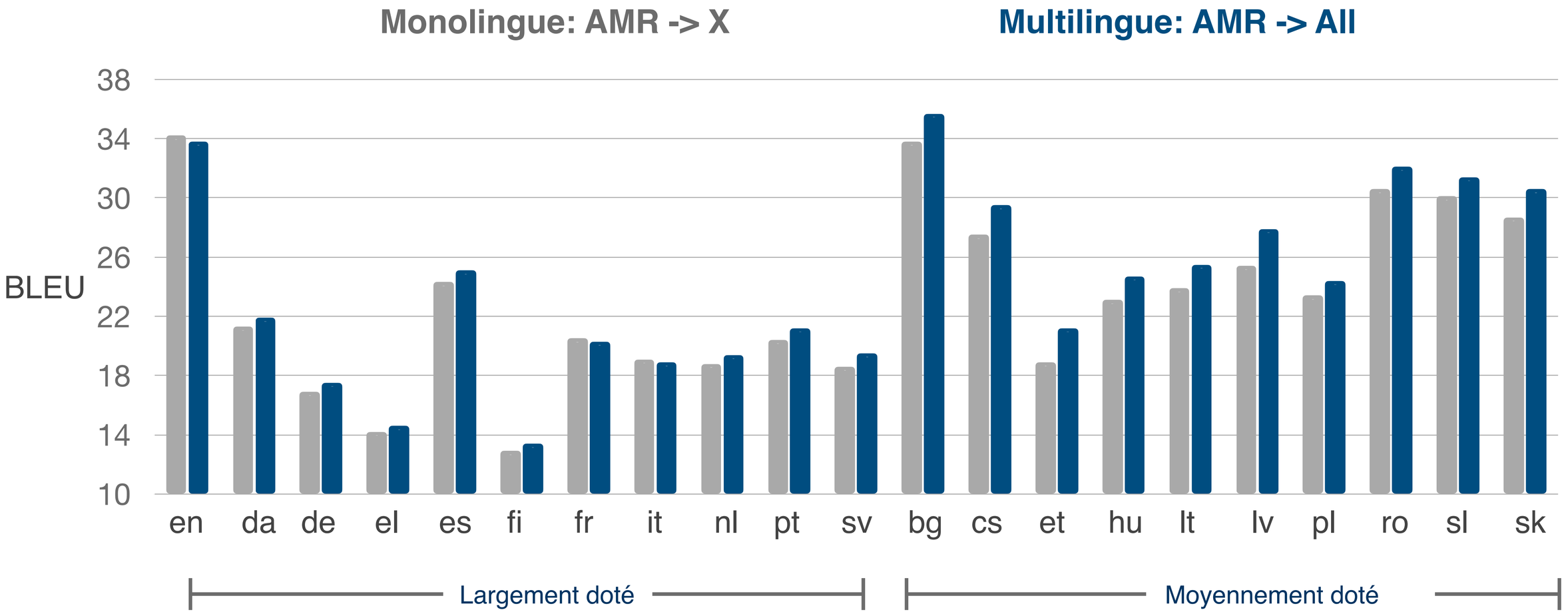
fr

Multilingue

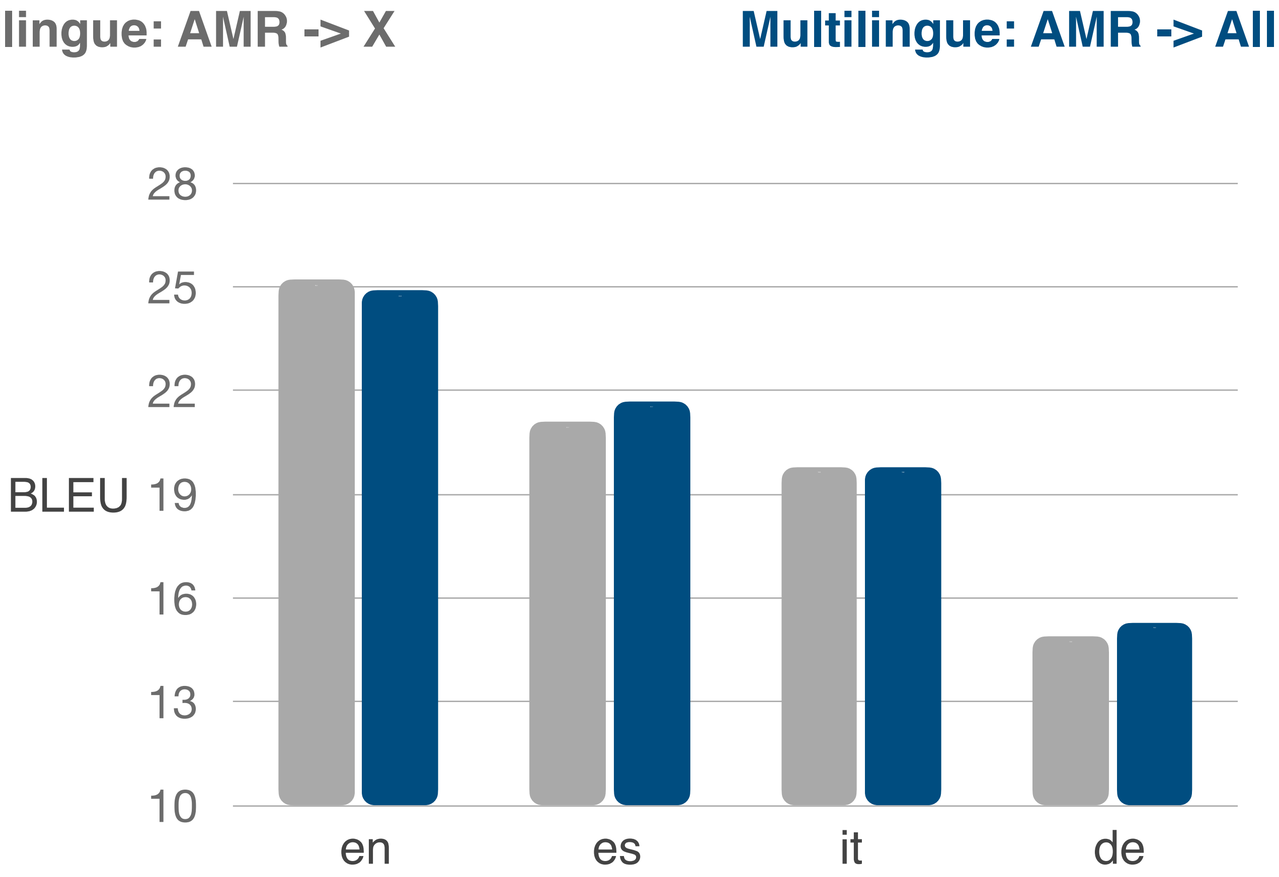
hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Résultats: AMR argent



Résultats: AMR or



Comparaison: NLG+Traduction vs. Bout-en-bout

Génération+Traduction

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓ AMR - Anglais

US officials held an expert group meeting in January 2002 in New York.

↓ Anglais - X

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Bout-en-bout multilingue
AMR-X

fr

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparaison : NLG+Traduction vs. Bout-en-bout

Génération+Traduction

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓ AMR - Anglais

US officials held an expert group meeting in January 2002 in New York.

↓ Anglais-X

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

fr

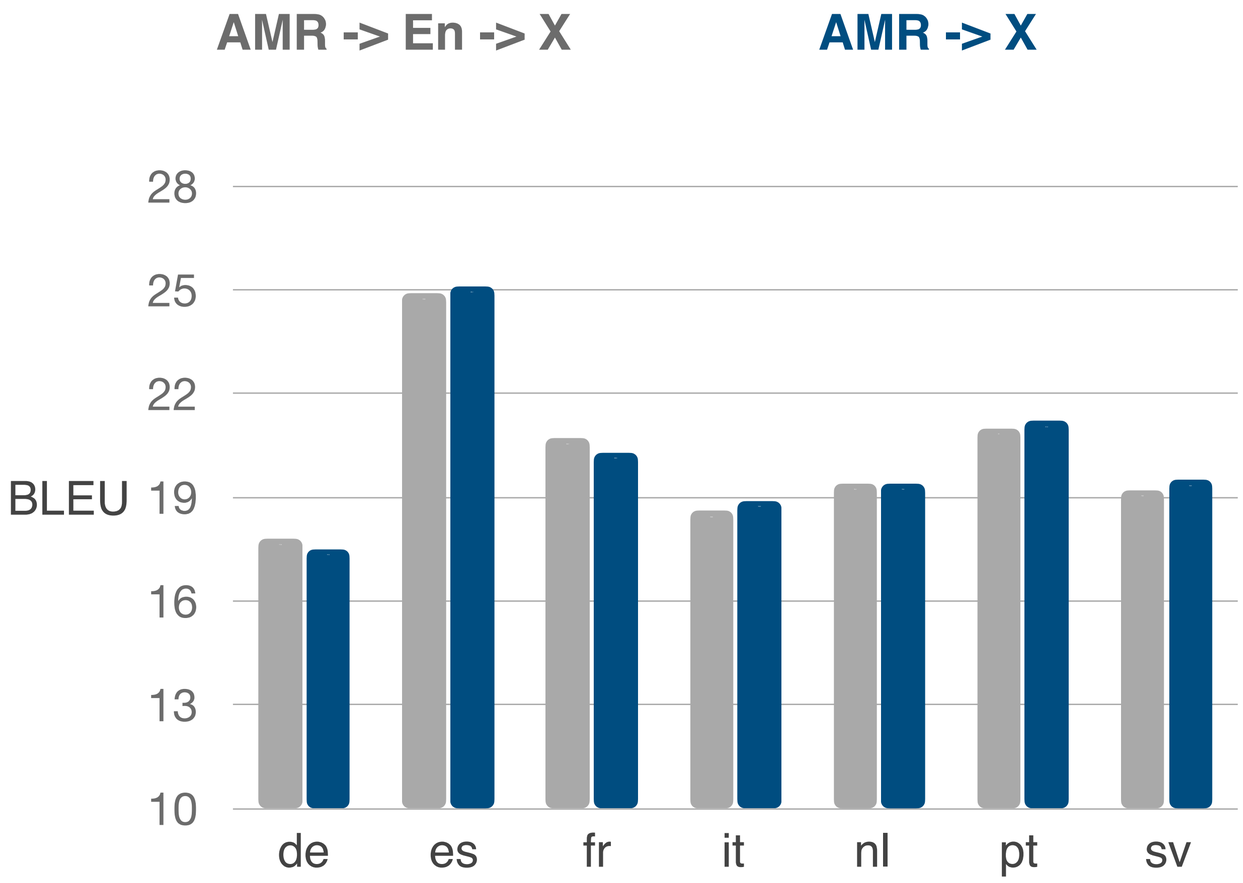
Bout-en-bout multilingue AMR-X

hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York

↓

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Comparaison : NLG+Traduction vs. Bout-en-bout

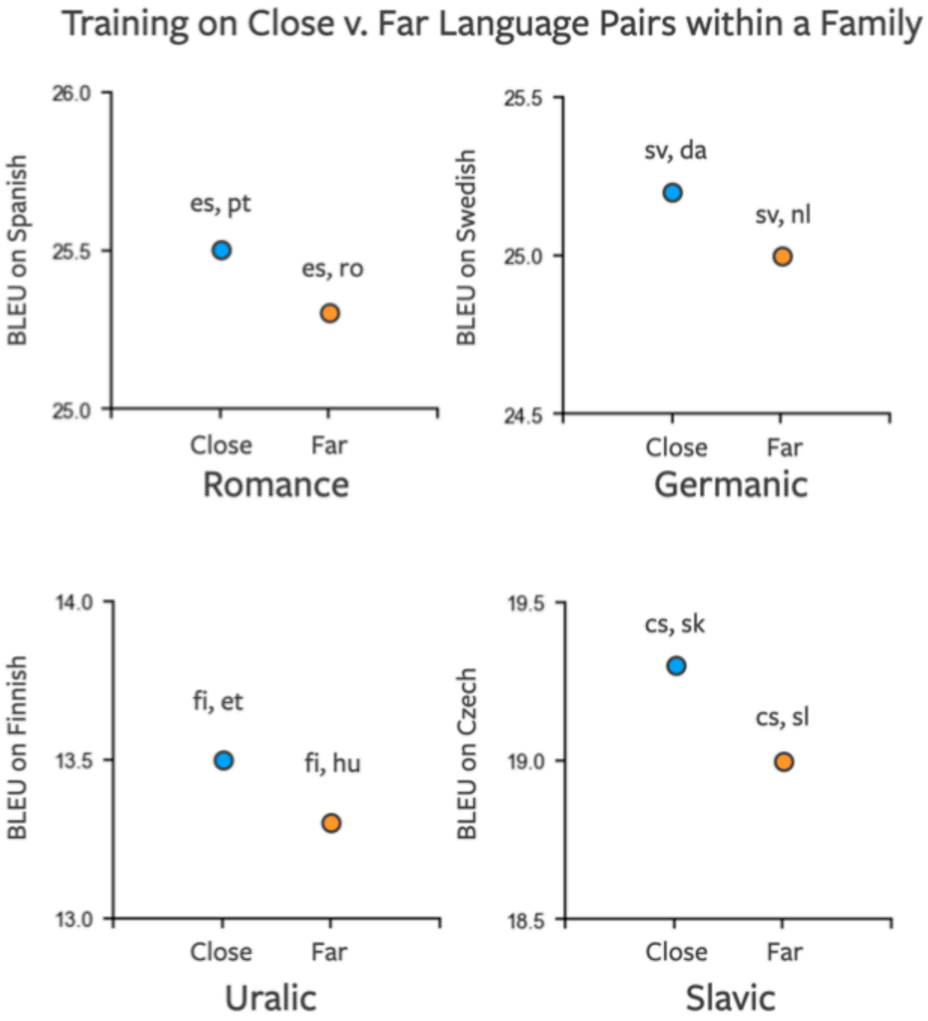


Impact des langues d'apprentissage

	Da	De	Nl	Sv
Une langue	21.3	17.0	18.5	18.7
Langues germaniques	21.8	21.9	19.6	19.3
21 langues	21.9	17.5	19.4	19.5

Apprendre d'une langue proche

- Modèles bilingues
- Pour une famille donnée, les paires de langues les plus proches donnent les meilleurs résultats
  - Romane: Espagnol/Portuguais
  - Germanique: Suédois/Danois
  - Uraliénne: Finnois/Estonien
  - Slaves : Tchèque/Slovaque
-



Evaluation par l'humain

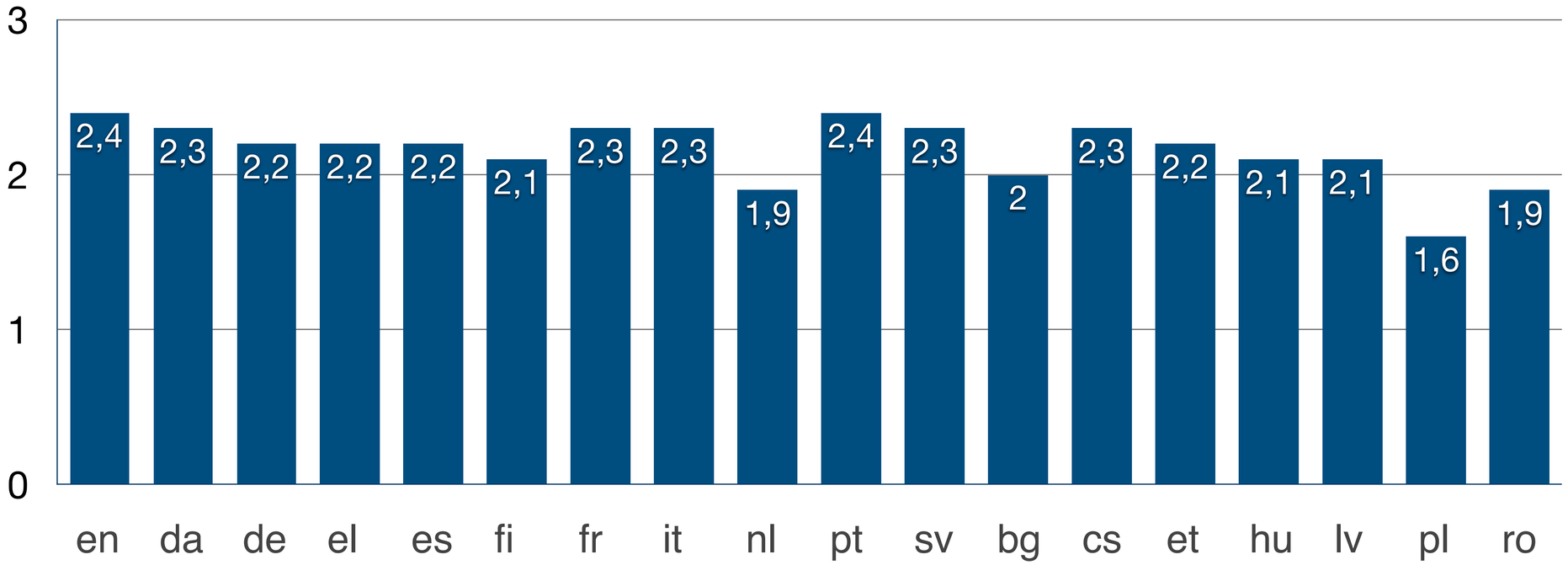
- **Sémantique :**
  - La phrase générée a-t-elle le même sens que la phrase de référence?

- **Morphologie:**
  - La morphologie est-elle correcte ? Les contraintes d'accord sont-elles respectées par exemple entre le verbe et le sujet ou entre le nom et l'adjectif?

- **Ordre des mots**
  - L'ordre des mots est-il naturel?

Sémantique

Un score de 2 indique des différences mineures



Paraphrases générées

REF: This point will certainly be the subject of subsequent further debates in the council

GEN: This is a point that will undoubtedly be discussed later in the council.

REF: Je ne suis pas favorable à des exceptions à cette règle.

GEN: A mon avis, il n'est pas bon de faire des exceptions à cette règle

.

Evaluation par l'humain

Les résultats sont bons également pour la morphologie et pour l'ordre des mots

Un modèle multilingue généralise bien à l'ensemble des langues étudiées

SY **You**

convert the amr written between "" to French "hold
:ARG0 person : ARG0-of have-org-role :ARG1 :op1 United :op2 States :ARG2 official
:ARG1 meet :ARG0 person :ARG1-of expert :ARG2-of group
:time date-entity :year 2002 :month 1
:location city :op1 New :op2 York"

ChatGPT

Organiser une réunion aux États-Unis à New York en janvier 2002. La réunion implique une personne occupant un rôle officiel et faisant également partie d'un groupe d'experts.

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

L'évaluation par l'humain montre que les techniques multilingues permettent de généraliser à 21 langues

L'évaluation par l'humain montre que les techniques multilingues permettent de généraliser à 21 langues

Le modèle multilingue bénéficie d'une plus grande quantité de données et dans de meilleurs résultats que les modèles monolingues

L'approche proposée permet de générer à partir d'une AMR anglo-centrique des textes dans 21 langues différentes

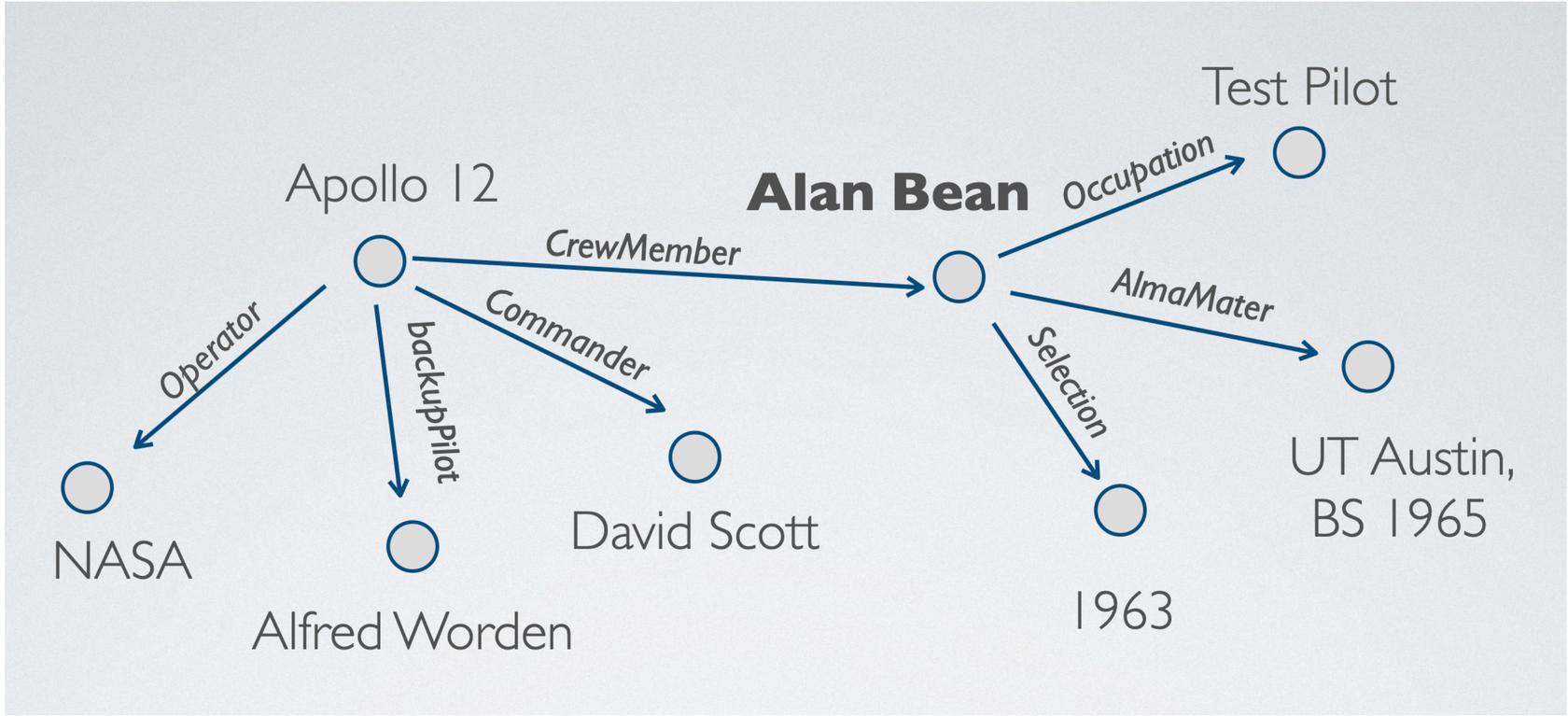
L'évaluation par l'humain indique que les techniques multilingues permettent de généraliser à 21 langues

Le modèle multilingue bénéficie d'une plus grande quantité de données et dans de meilleurs résultats que les modèles monolingues

Générer à partir de graphes de connaissances

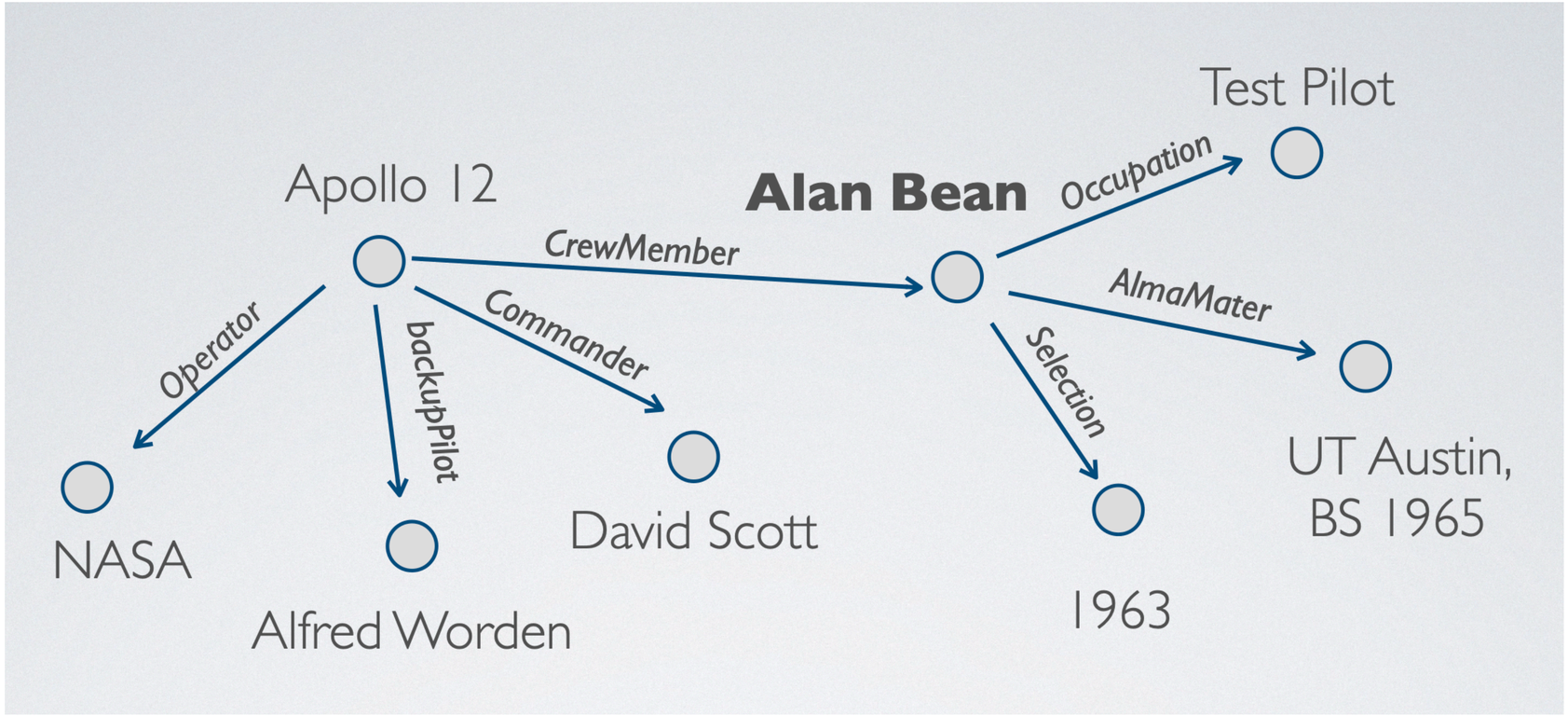
Gardent et al. ACL 2017, Castro-Ferreira et al. 2020, Cripwell et al. 2023
W. Soto-Martinez, Y. Parmentier and C. Gardent AACL 2023

Défi partagé WebNLG



Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred. Worden and was commanded by David Scott

Défi partagé WebNLG



Gardent et al. ACL 2017, Castro-Ferreira et al. 2020, Cripwell et al. 2023

WebNLG 2017 : RDF ==> Anglais

	Train+Dev	Test (Seen Category)	Test (Unseen Category)	TOTAL
# (Graph,Text)	20,370	2,495	2,413	**25,298**
# Graphs	7,812	971	891	**9,674**

- Graphes DBPedia avec des entités racines de différentes catégories

- Textes écrits par des humains

Gardent et al. ACL 2017

WebNLG 2017 : RDF ==> Anglais

	Train+Dev	Test (Seen Category)	Test (Unseen Category)	TOTAL
# (Graph,Text)	20,370	2,495	2,413	**25,298**
# Graphs	7,812	971	891	**9,674**

10 catégories connues

- Astronaute, Université, Monument, Construction, Comique, Aliment, Aéroport, EquipeSportive, Ville, Ecrit

5 catégories inconnues

- Athlète, Artiste, MoyenDeTransport, CorpsCéleste, Politique

Gardent et al. ACL 2017

WebNLG 2017 : RDF ==> Anglais

6 participants, 10 systèmes

- 3 modèles symboliques
- 1 modèle statistique
- 5 modèles neuronaux

Gardent et al. ACL 2017

WebNLG 2017 : RDF ==> Anglais

Tout

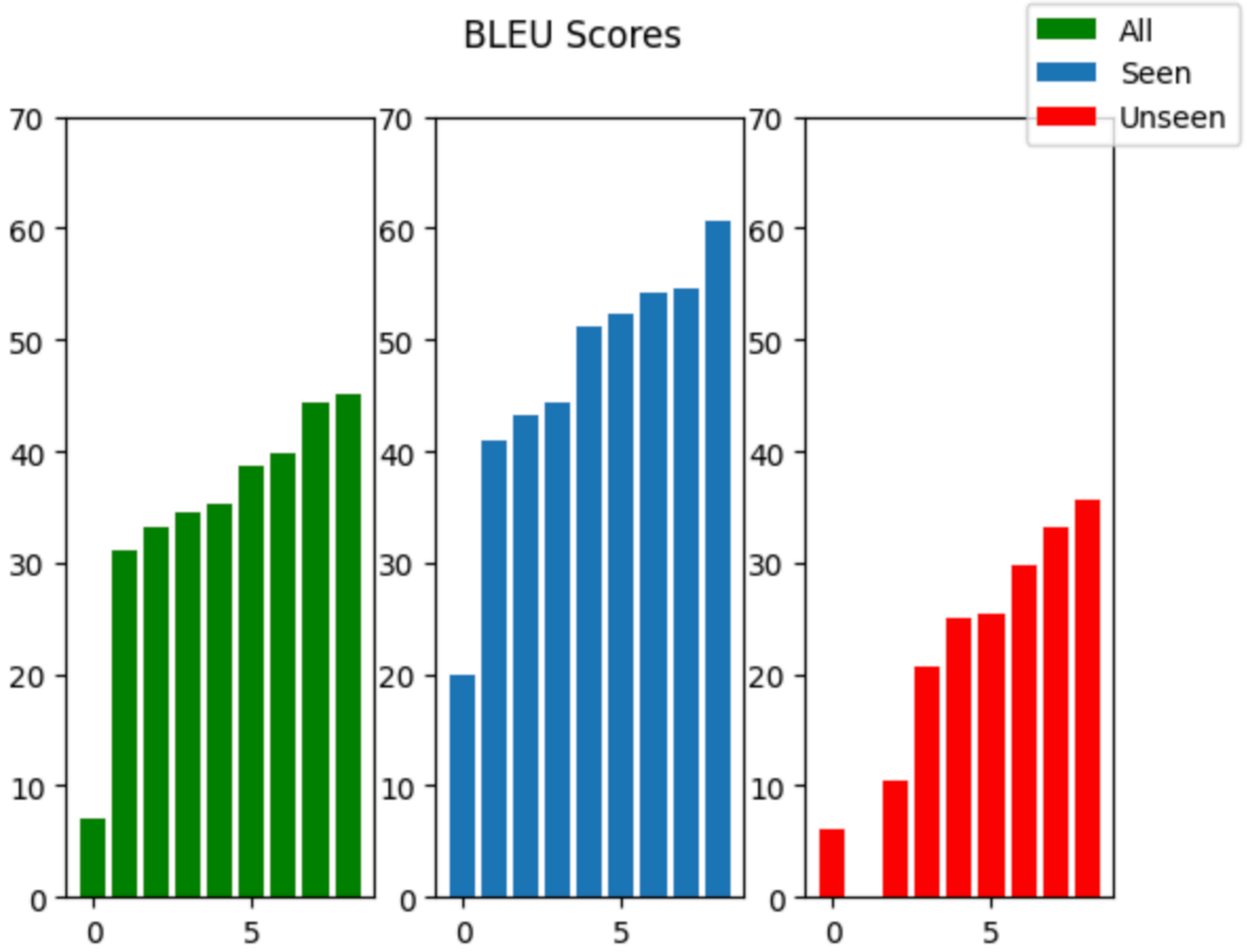
7.07 - 45.13

Connu

19.87 - 60.54

Inconnu

5.13 - 35.7



Gardent et al. ACL 2017

WebNLG 2020

Génération

- RDF ==> Anglais

Castro-Ferreira et al. INLG 2020

WebNLG 2020

Génération

- RDF ==> Anglais
- RDF ==> Russe

Castro-Ferreira et al. INLG 2020

WebNLG 2020

Génération

- RDF ==> Anglais
- RDF ==> Russe

Analyse

- Anglais ==> RDF
- Russe ==> RDF

Castro-Ferreira et al. INLG 2020

WebNLG 2020

	Train	Dev	Test NLG/SP	TOTAL
# (Graph,Text)	35,426	4,664	5,150	47,395
# Graphs	13,211	1,667	1,779	17,409

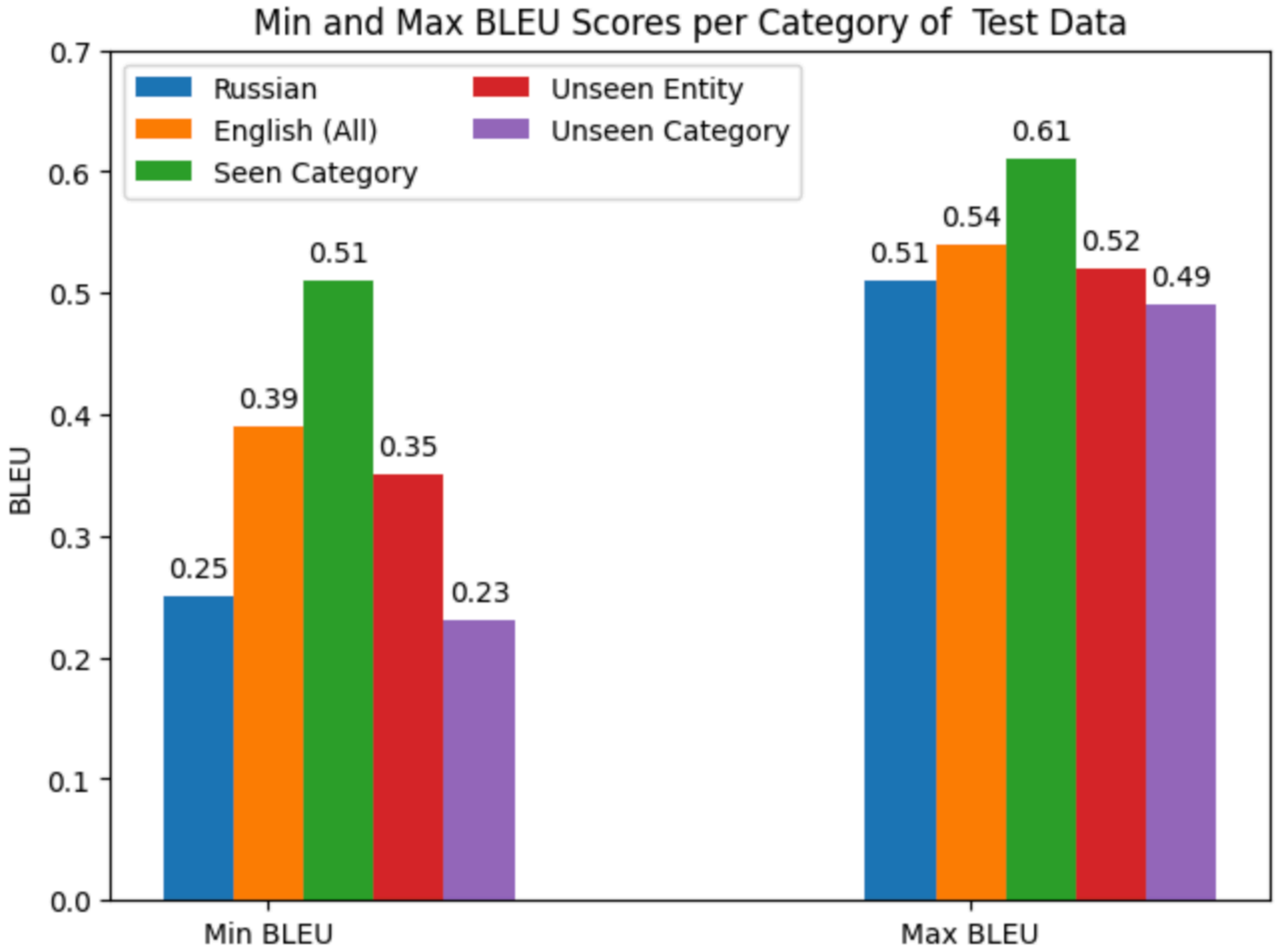
- 16 catégories connues
- 3 catégories inconnues
- Entités inconnues

WebNLG 2020 : Participation

System	Affiliation	Country
MED	Sber AI Lab	Russia
RALI-UMONTRÉAL	Université de Montréal	Canada
ORANGE-NLG	Orange Labs	France
CUNI-UFAL	Charles University	Czechia
TGEN	AIST	Japan
BT5	Google	US
UPC-POE	Universitat Politècnica de Catalunya	Spain
DANGNT-SGU	Saigon University	Vietnam
HUAWEI	Huawei Noah's Ark Lab	UK
AMAZONAI	Amazon AI (Shanghai)	China
NILC	University of São Paulo	Brazil
NUIG-DSI	National University of Ireland	Ireland
CYCLEGT	Amazon	China
OSU NEURAL NLG	The Ohio State University	US
FBCONVAI	Facebook	US

17 participants

WebNLG 2020 : Résultats



WebNLG 2023 : Langues peu dotées

	Silver Train	Dev	Test
Breton	13,211	1,399	1,778
Welsh	13,211	1,665	1,778
Irish	13,211	1,665	1,778
Maltese	13,211	1,665	1,778

WebNLG 2023 : Génération + Traduction

Team	Affiliation	Country	Breton	Welsh	Irish	Maltese	Russian
CUNI-Wue	Charles University	Czechia	✓	✓	✓	✓	✓
DCU/TCD-FORGe	ADAPT/DCU/Trinity College	Ireland	-	-	✓	-	-
Interno	Pulkovo Observatory	Russia	-	-	-	-	✓
IREL	IIT Hyderabad	India		✓	✓	✓	✓
DCU-NLG-PBN	ADAPT/DCU	Ireland	-	✓	✓	✓	-

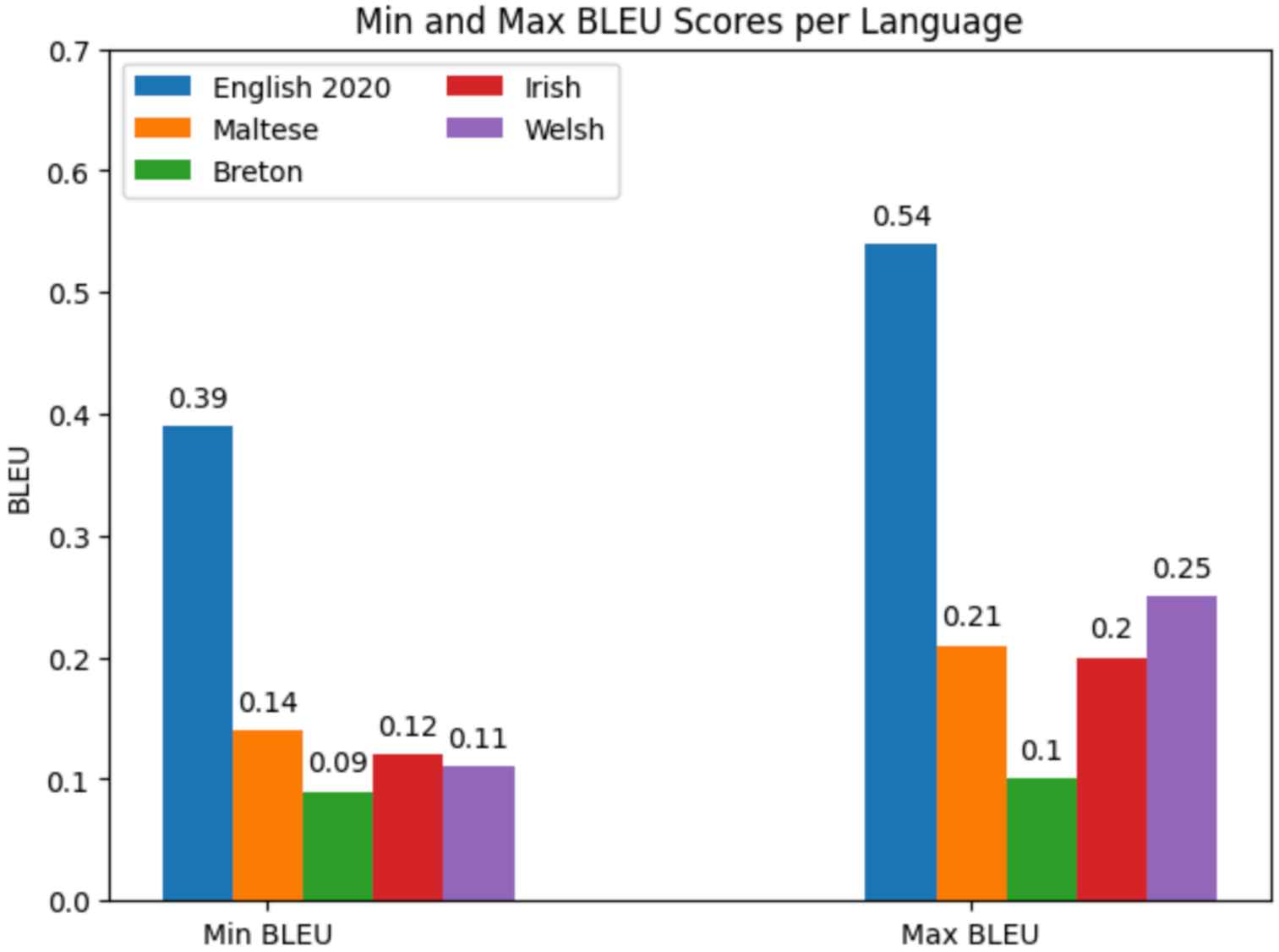
RDF ==> Anglais

- T5 ou mT5 affiné sur les données WebNLG (anglais)
- GPT3-5, in context learning

Anglais ==> Langue peu dotée

- Traduction automatique : NLLB ou Google Translate

WebNLG 2023 : Résultats



Modèle en cascade vs. De bout en bout

Pour le breton, la traduction automatique est mauvaise

❌ Génération + Traduction

Modèle en cascade vs. De bout en bout

Pour le breton, la traduction automatique est mauvaise

- ✗ Génération + Traduction
- ✗ Affinage (BLEU : 0.10)

Modèle en cascade vs. De bout en bout

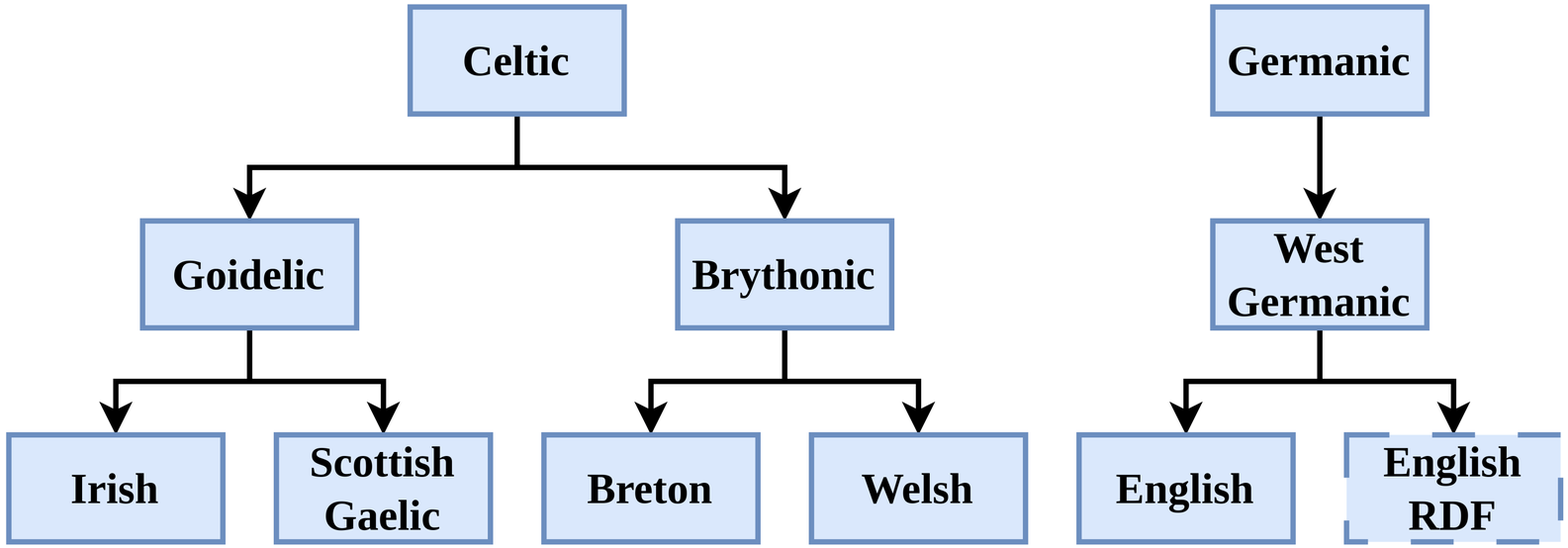
Pour le breton, la traduction automatique est mauvaise

Génération + Traduction

Affinage (BLEU : 0.10)

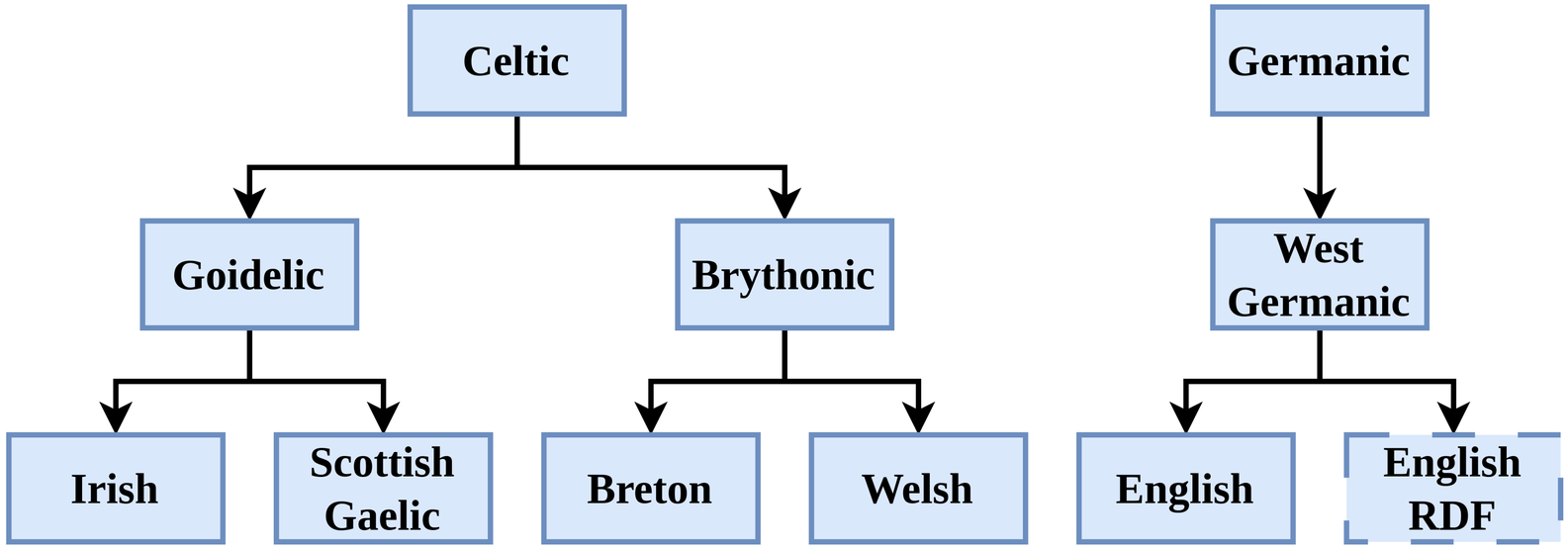
Affinage + Préfixe continu

WebNLG 2023 : Préfixe continu



- Préfix continu
- Structuré pour modéliser les relations entre langues

WebNLG 2023 : Préfixe continu



Préfix continu

50 Tokens Task	15 Tokens Source Family	15 Tokens Source Genus	15 Tokens Source Language	15 Tokens Target Family	15 Tokens Target Genus	15 Tokens Target Language	n Tokens Input Sequence

Apprentissage

Etape 1. Apprentissage auto-supervisé (Modèles de langue)

Le préfixe continu est appris sur des tâches monolingues

Input Batch

Task	Source			Target			Original Input Sequences					
	Family	Genus	Lang.	Family	Genus	Lang.						
Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Apprentissage

Etape 1. Apprentissage auto-supervisé (Modèles de langue)

Le préfixe continu est appris sur des tâches monolingues

Input Batch

Task	Source			Target			Original Input Sequences					
	Family	Genus	Lang.	Family	Genus	Lang.						
Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Etape 2. Affinage sur les données RDF-Texte

Apprentissage

Etape 1. Apprentissage auto-supervisé (Modèles de langue)

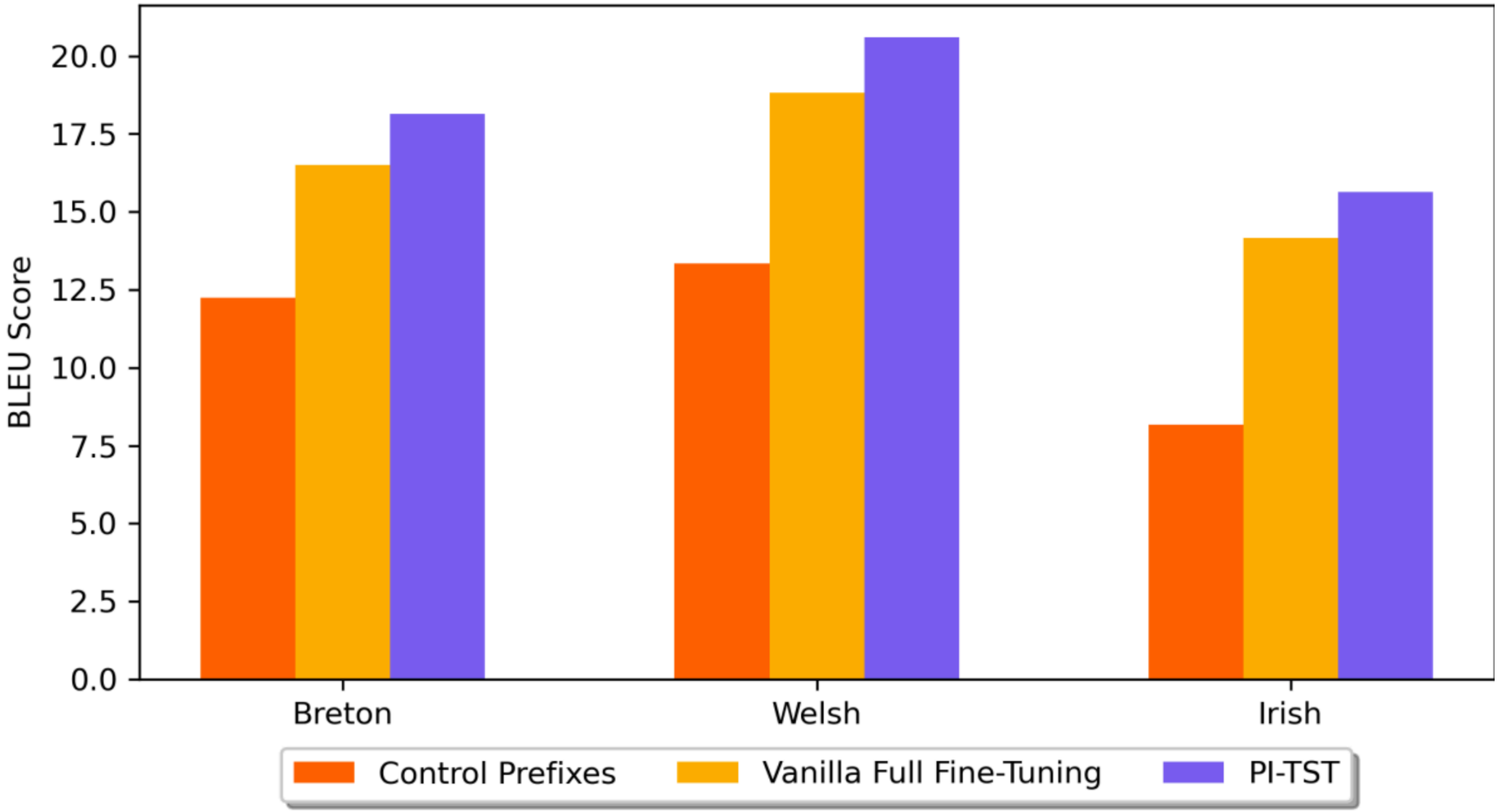
Le préfixe continu est appris sur des tâches monolingues

Input Batch	Task	Source			Target			Original Input Sequences					
		Family	Genus	Lang.	Family	Genus	Lang.						
	Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
	Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
	Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
	Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
	Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Etape 2. Affinage sur les données RDF-Texte

Inférence. On utilise le préfixe de la langue cible

WebNLG 2023 : Résultats



Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)

Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)
- Les résultats sont moins bons pour le russe que pour l’anglais

Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)
- Les résultats sont moins bons pour le russe que pour l’anglais
- Les performances se dégradent sur les données hors-domaines (inconnues)

Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)
- Les résultats sont moins bons pour le russe que pour l’anglais
- Les performances se dégradent sur les données hors-domaines (inconnues)
- Les résultats sont mauvais pour les langues peu dotées

Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)
- Les résultats sont moins bons pour le russe que pour l’anglais
- Les performances se dégradent sur les données hors-domaines (inconnues)
- Les résultats sont mauvais pour les langues peu dotées
- Les modèles en cascade proposés pour ces langues fonctionnent mal lorsque la traduction automatique est mauvaise

Points clés

- Le pré-apprentissage améliore les résultats (2017 vs. 2020)
- Les résultats sont moins bons pour le russe que pour l'anglais
- Les performances se dégradent sur les données hors-domaines (inconnues)
- Les résultats sont mauvais pour les langues peu dotées
- Les modèles en cascade proposés pour ces langues fonctionnent mal lorsque la traduction automatique est mauvaise
- Un préfixe continu permet d'améliorer ces résultats
  BLEU breton : 10 (NLG+MT) —> 18.15 (PEFT)

Générer à partir de textes

Génération de biographies

Angela **Fan and Claire Gardent**
“Generating Full Length Wikipedia Biographies. The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies.”
ACL 2022

Générer des biographies Wikipédia à partir du Web

PERSON NAME

Modèle RAG

Génération + Extraction d'information

WIKIPEDIA

Joan Paton

Joan Burton Paton AM née Cleland (1916–April 2000) was an Australian teacher, naturalist, environmentalist and ornithologist. One of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected President 1991–1993, she also served as president of the South Australian Ornithological Association (1979-1982). Her father was Professor Sir John Burton Cleland, a notable microbiologist and pathologist who strongly encouraged her early interest in natural history.

Contents

Early life and education

Joan Burton Paton was born in Sydney, New South Wales, the daughter of John Burton Cleland (1878–1971) and his wife, Dora Isabel Paton (1880–1955).[1] She had three sisters, Dr Margaret Burton Cleland, Elizabeth Robson Cleland and Barbara Burton Cleland; and a brother, William Paton 'Bill' Cleland, who became a surgeon. The father encouraged his children's interest in science. Joan Paton was educated at the University of Adelaide, where she majored in organic chemistry and biochemistry. In 1951 she married Erskine Norman Paton (1922–1985), son of Adolph Ernest Paton and Ida Marie Poynton. Their son is Prof David Cleland Paton.[2]

Career

In 1967 Paton became a lecturer on ornithology in South Australia's Workers' Educational Association.[3][4] Among those she inspired to work in ornithology and environmental conservation was Margaret Cameron, who became the President of the Royal Australasian Ornithologists Union (RAOU).[5]

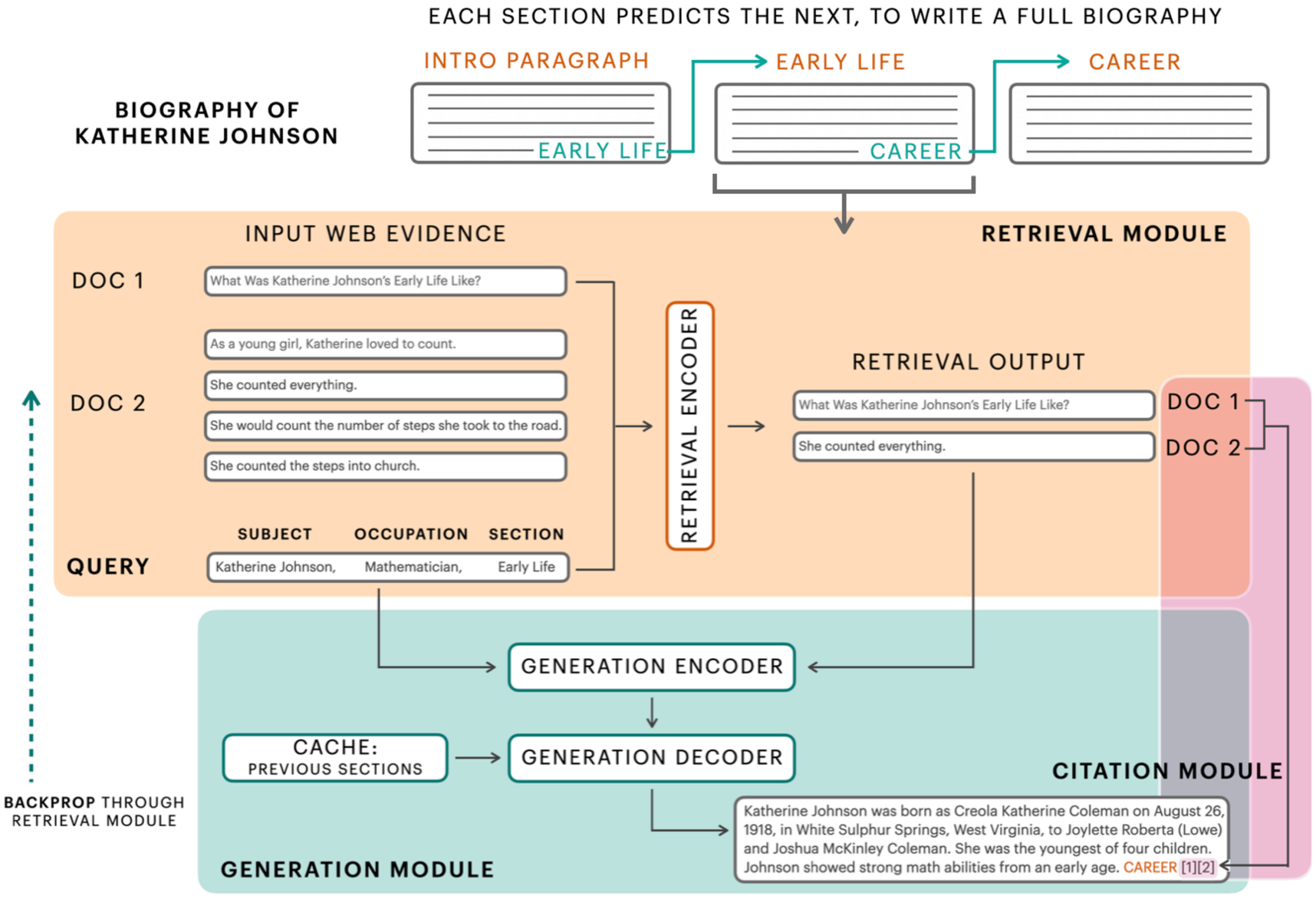
Paton was active in the RAOU, as well as in the South Australian Ornithological Association (SAOA), of which she was elected Vice-President 1974–1979, and President 1979–1982. She was one of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected president (1991-1993).[6]

Legacy and honours

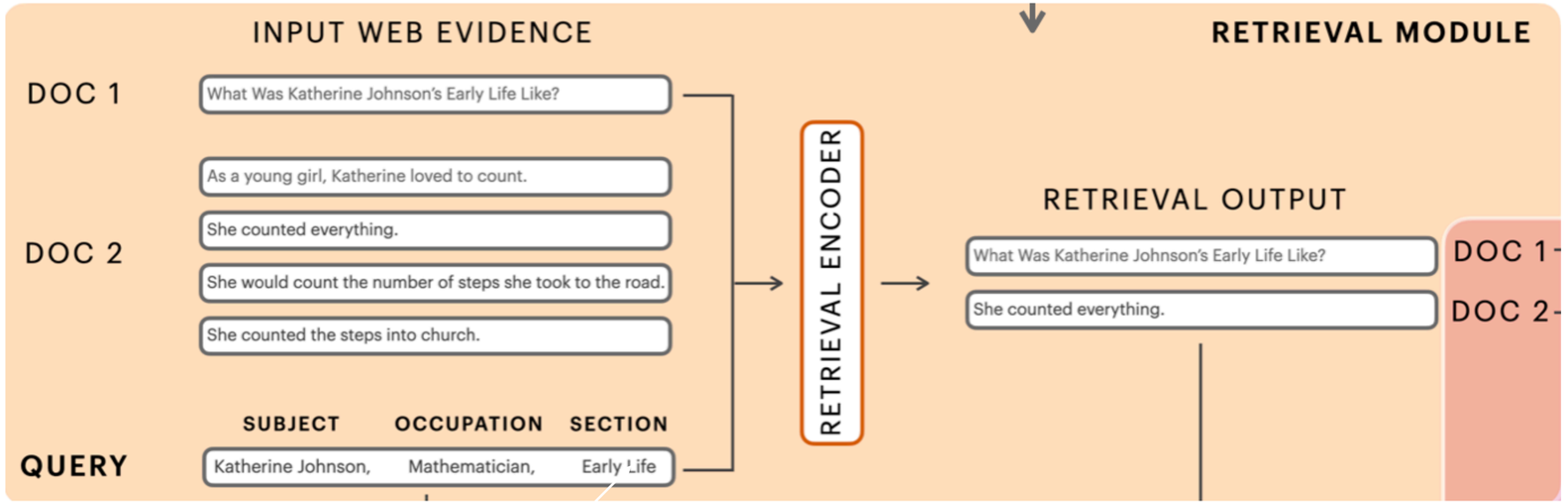
- 1990, she was made an Honorary Member of the SAOA.
- 1996, she was made an Honorary Member of the Adelaide Ornithologists Club.

Enjeux

- Rassembler des informations pertinentes
- Produire un texte structuré
- S'assurer de la véracité des faits


EACH SECTION PREDICTS THE NEXT, TO WRITE A FULL BIOGRAPHY
BIOGRAPHY OF KATHERINE JOHNSON
INTRO PARAGRAPH
EARLY LIFE
CAREER
EARLY LIFE
CAREER
INPUT WEB EVIDENCE
RETRIEVAL MODULE
DOC 1
What Was Katherine Johnson's Early Life Like?
DOC 2
As a young girl, Katherine loved to count.
She counted everything.
She would count the number of steps she took to the road.
She counted the steps into church.
QUERY
SUBJECT
OCCUPATION
SECTION
Katherine Johnson,
Mathematician,
Early Life
RETRIEVAL ENCODER
RETRIEVAL OUTPUT
What Was Katherine Johnson's Early Life Like?
She counted everything.
DOC 1
DOC 2
GENERATION ENCODER
CACHE:
PREVIOUS SECTIONS
GENERATION DECODER
CITATION MODULE
Katherine Johnson was born as Creola Katherine Coleman on August 26, 1918, in White Sulphur Springs, West Virginia, to Joylette Roberta (Lowe) and Joshua McKinley Coleman. She was the youngest of four children. Johnson showed strong math abilities from an early age. CAREER [1][2]
GENERATION MODULE
BACKPROP THROUGH RETRIEVAL MODULE

Extraction



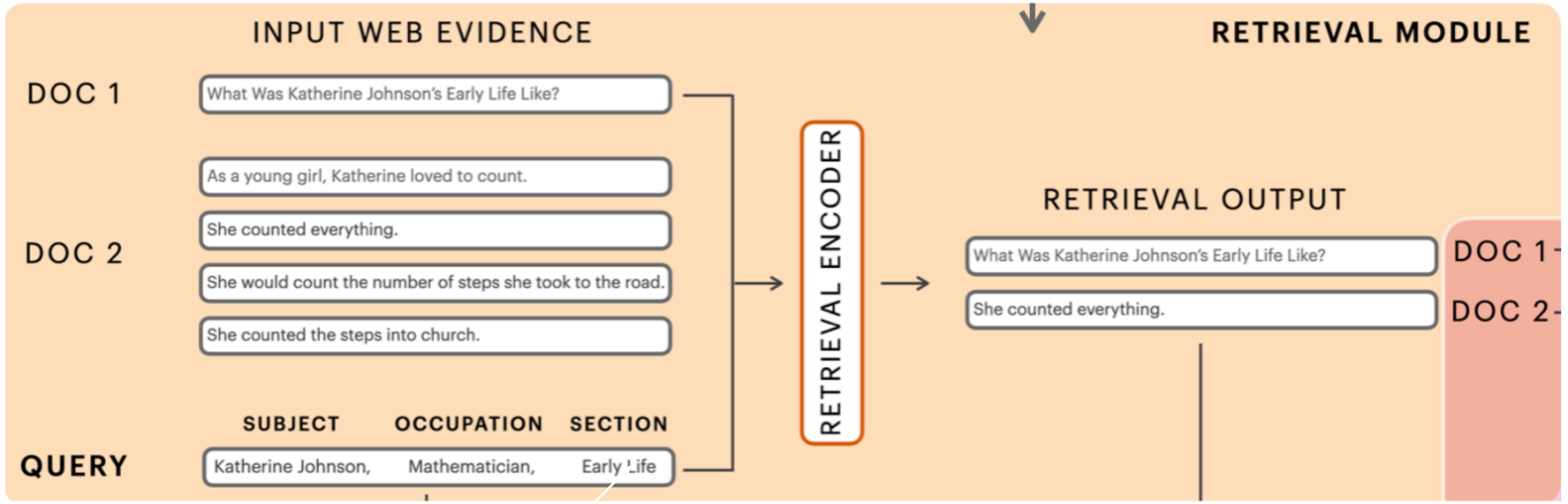
REQUÊTE

Katherine Johnson

Mathematician

Early Life

Extraction



REQUÊTE

Katherine Johnson

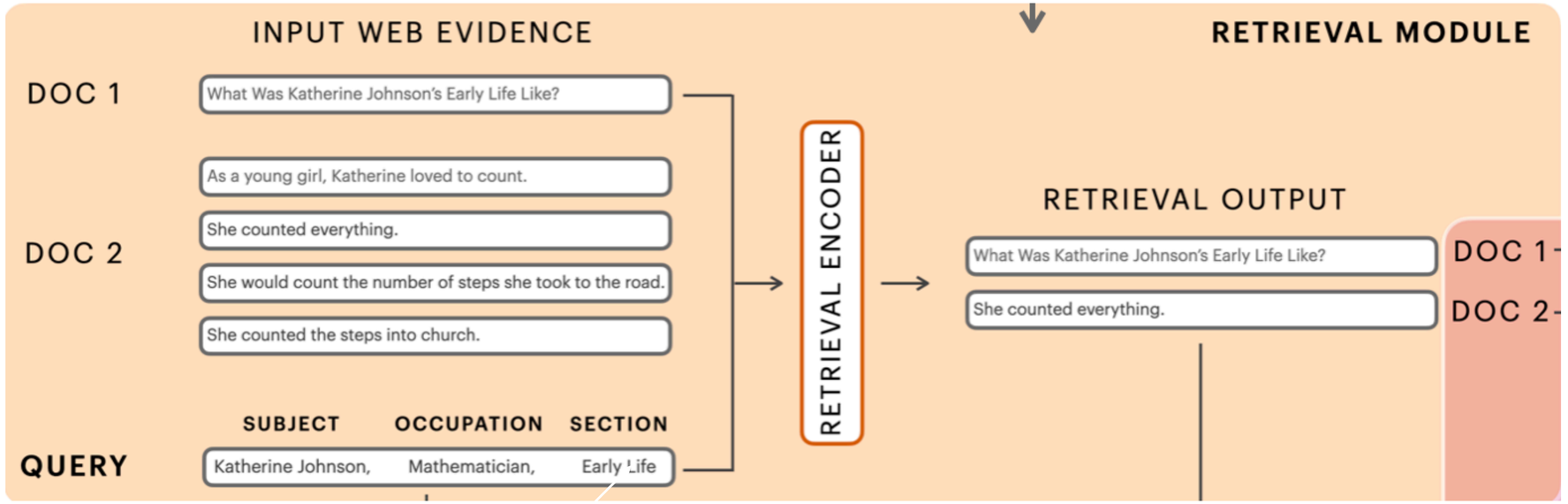
Mathematician

Early Life

RESULTAT

20 premiers documents segmentés en phrase

Extraction



REQUÊTE

Katherine Johnson

Mathematician

Early Life

RESULTAT

20 premiers documents segmentés en phrase

SORTIE

40 phrases les plus similaires à la requête

(1,000 mots)

Génération

REQUÊTE

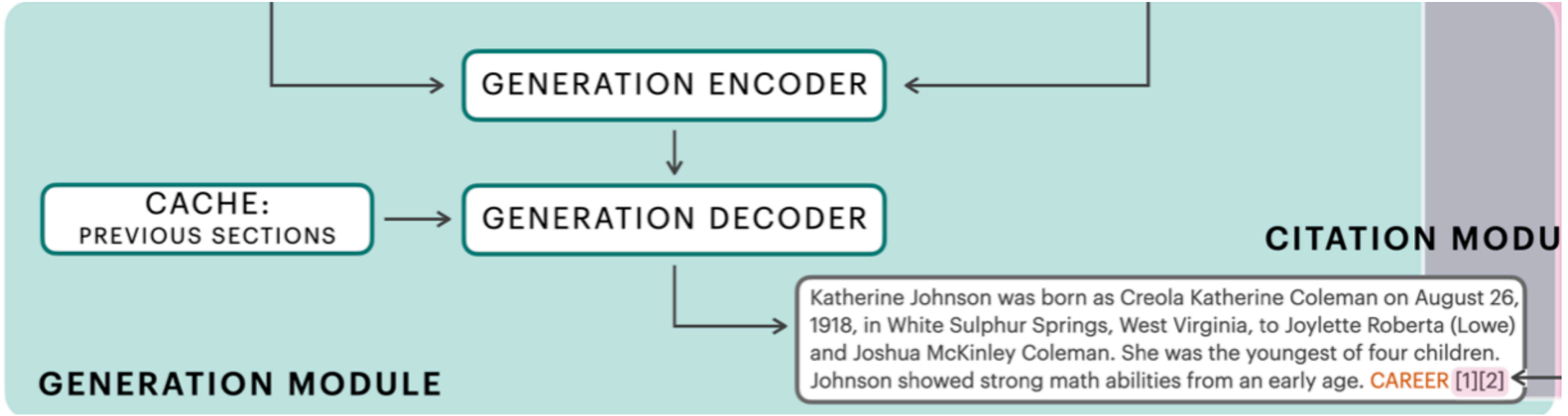
Katherine Johnson

Mathematician

Early Life

EXTRACTION

1,000 mots



Cache Transformer-XL

- Les états cachés des sections précédentes sont conservés
- Servent de mémoire pour la génération de la section suivante

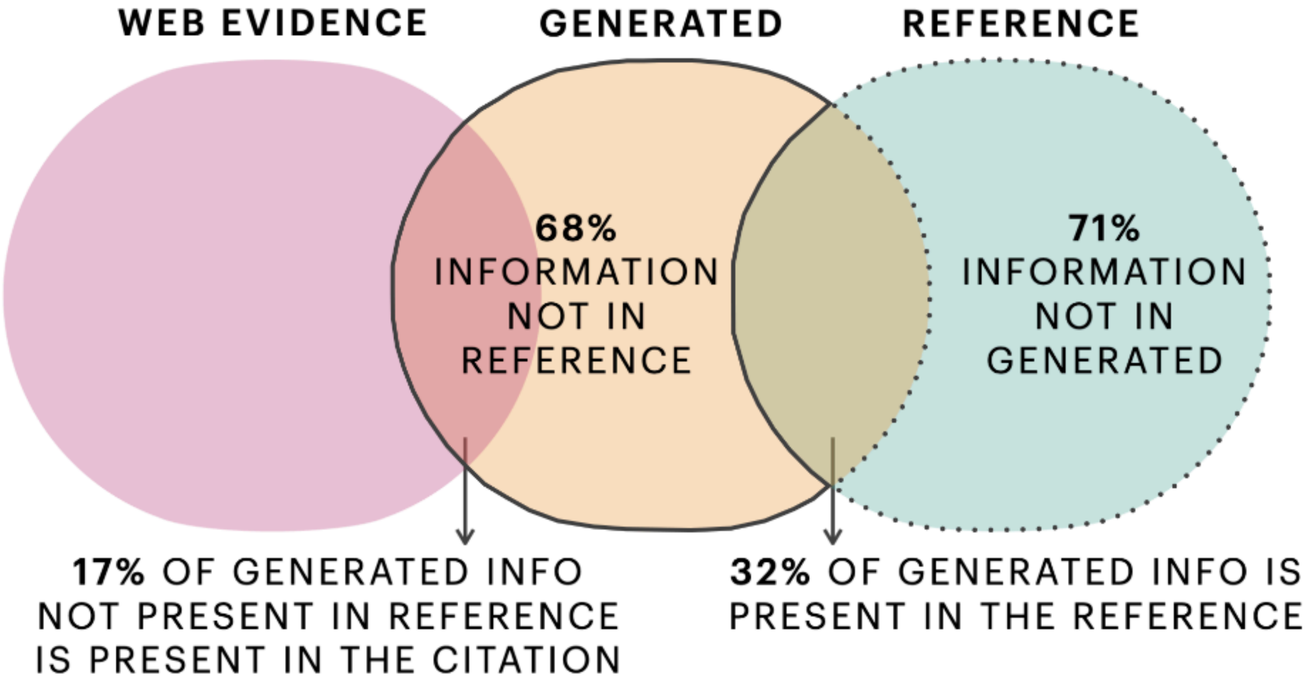
Ablation

Model	ROUGE-L	Entailment	Named Entity Coverage
BART Pretraining + Finetuning	17.4	15.8	21.9
+ Retrieval Module	18.8	17.2	23.1
+ Caching Mechanism	19.3	17.9	23.4

L'extraction et le module de cache permettent une amélioration des résultats qui est statistiquement significative

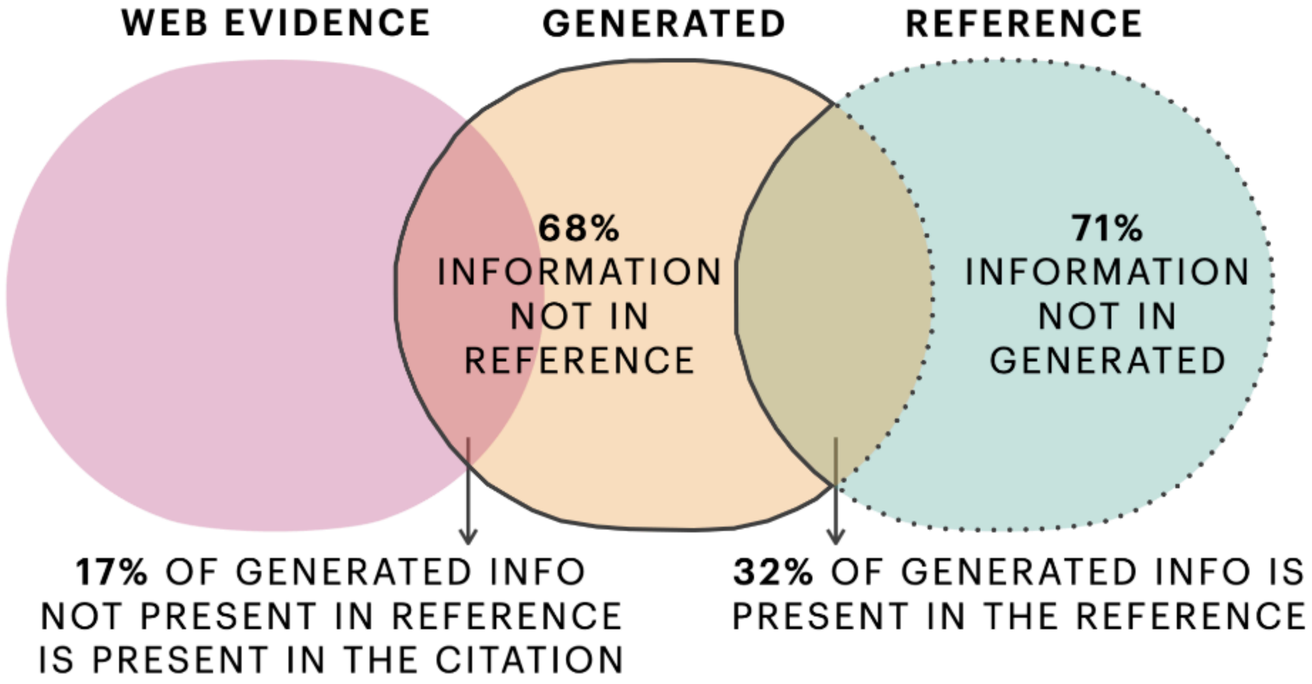
Evaluation par l'humain de la factualité

- Les textes générés ne contiennent qu'une faible partie des informations contenues dans le texte de référence



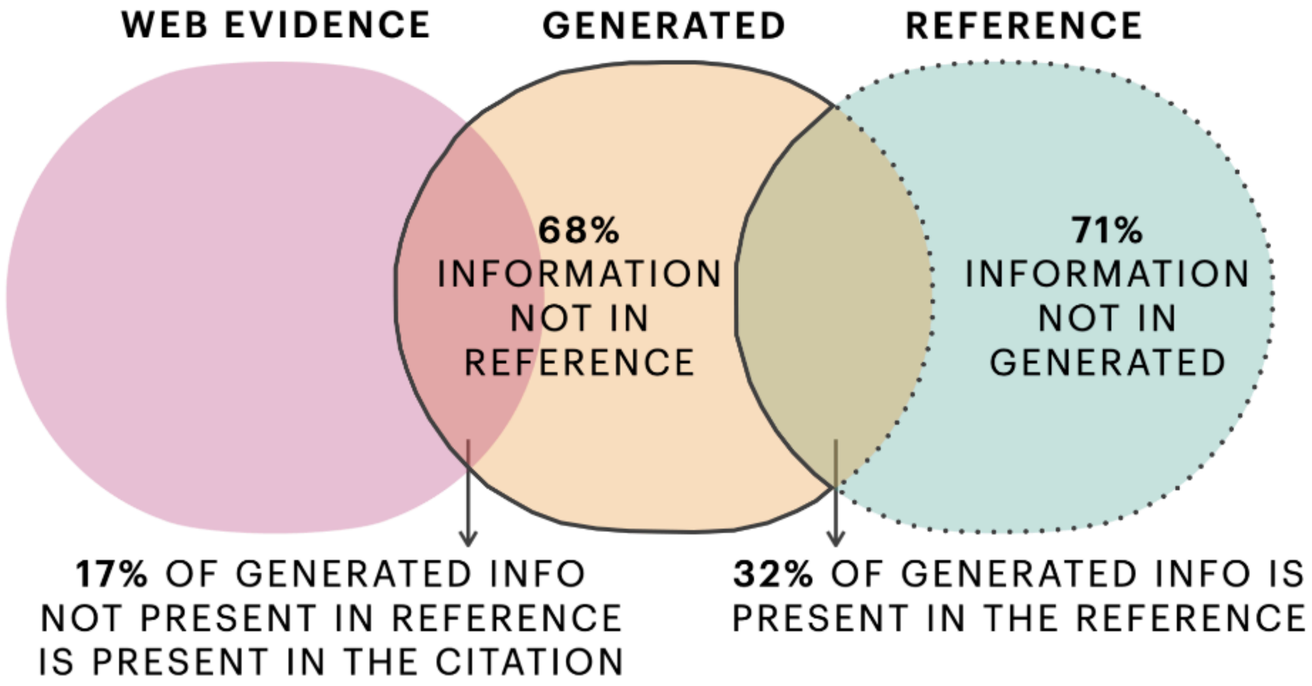
Evaluation par l'humain de la factualité

- Les textes générés ne contiennent qu'une faible partie des informations contenues dans le texte de référence
- La référence ne contient pas toutes les informations extraites de Web



Evaluation par l'humain de la factualité

- Les textes générés ne contiennent qu'une faible partie des informations contenues dans le texte de référence
- La référence ne contient pas toutes les informations extraites de Web
- Environ 50 % des informations contenues dans le texte généré ne sont pas avérées



Biais dans les données

Wikisum: biographies
Wikipedia

Our dataset: biographies
Wikipedia **de femmes**

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Biais dans les données

Les biographies Wikipedia **de femme** sont plus courtes

- Moins de sections
- Sections plus courtes
- Moins de mots

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Biais dans les données

Le web contient peu de documents sur les femmes

- En moyenne, pour les femmes, une requête Web ne retourne que 18 documents

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Biais dans les données

Le web contient peu d'information sur les femmes

- Les biographies Wikipedia de femmes ont une intersection plus faible avec les données extraites du Web

WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Moins d'information, une génération moins bonne

Model	WikiSum Test	Women	Scientists	Women in Asia	Women in Africa
BART Pretraining	19.0	17.4	18.2	16.7	16.4
+ Retrieval	21.4	18.8	19.3	17.9	17.1
+ Caching	21.8	19.3	19.7	18.4	17.3

- Les performances sont moins bonnes pour les biographies de femmes particulièrement les biographies de femmes asiatiques et africaines

Exemple de texte généré

---

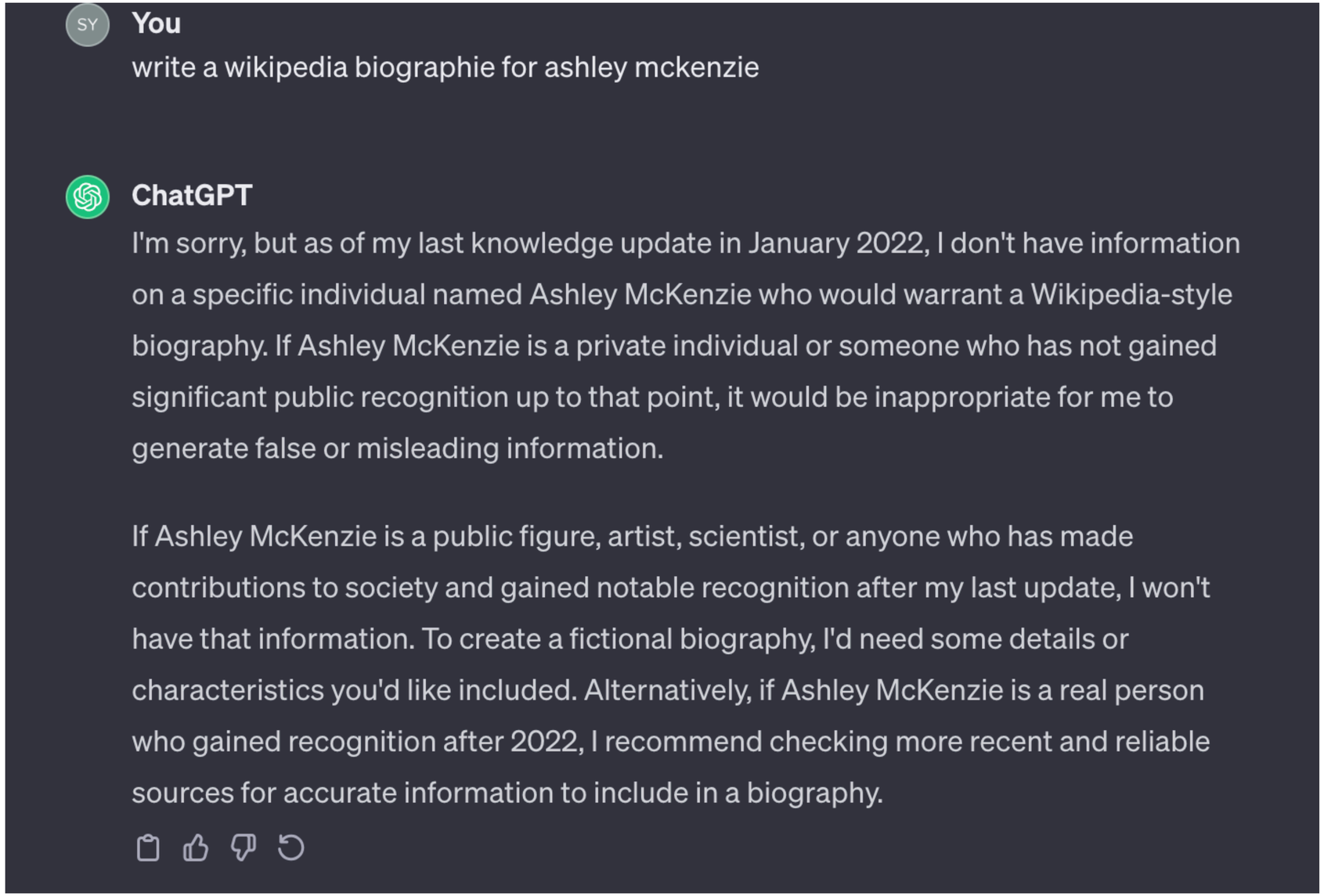
toplevel ashley mckenzie is a canadian film director, screenwriter and producer. she is the winner of the stella artois jay scott prize for emerging talent at the 2016 toronto international film festival. her first feature film, werewolf, premiered at the toronto film festival in 2016. she has also directed short films for the national film board of canada and the canadian screen actors guild. she was born in montreal, quebec, canada, and grew up in ottawa, ontario. [1,3,11,13,14]
=personal life= mckenzie was born in london, england. she is the daughter of alexander mckenzie, who was a member of the british rock band the beatles. she has a younger sister, jessica, who is also a singer. she was educated at st mary ’s college, oxford, where she graduated with a bachelor of arts degree in english literature. she also studied at the university of london. she married fellow x factor contestant andrew davies in september 2006. they have two children, a son and a daughter. [3,4,7,8,10,11]
=career= mckenzie was a contestant on the third series of the x - factor in 2006. she was eliminated in the first week of the competition. in 2007, mckenzie released her debut single "don ’t pretend you hadn’ t, now..." which peaked at no .160; 2 on the uk singles chart. she also released a second single ," i ’m not afraid ", in 2008. in 2009, she released her third single ," don’ t pretend you haven ’t, now ". in 2010, she was a judge on the x factor uk. [2]

ChatGPT 3.5
29/12/2023

ChatGPT ne connait

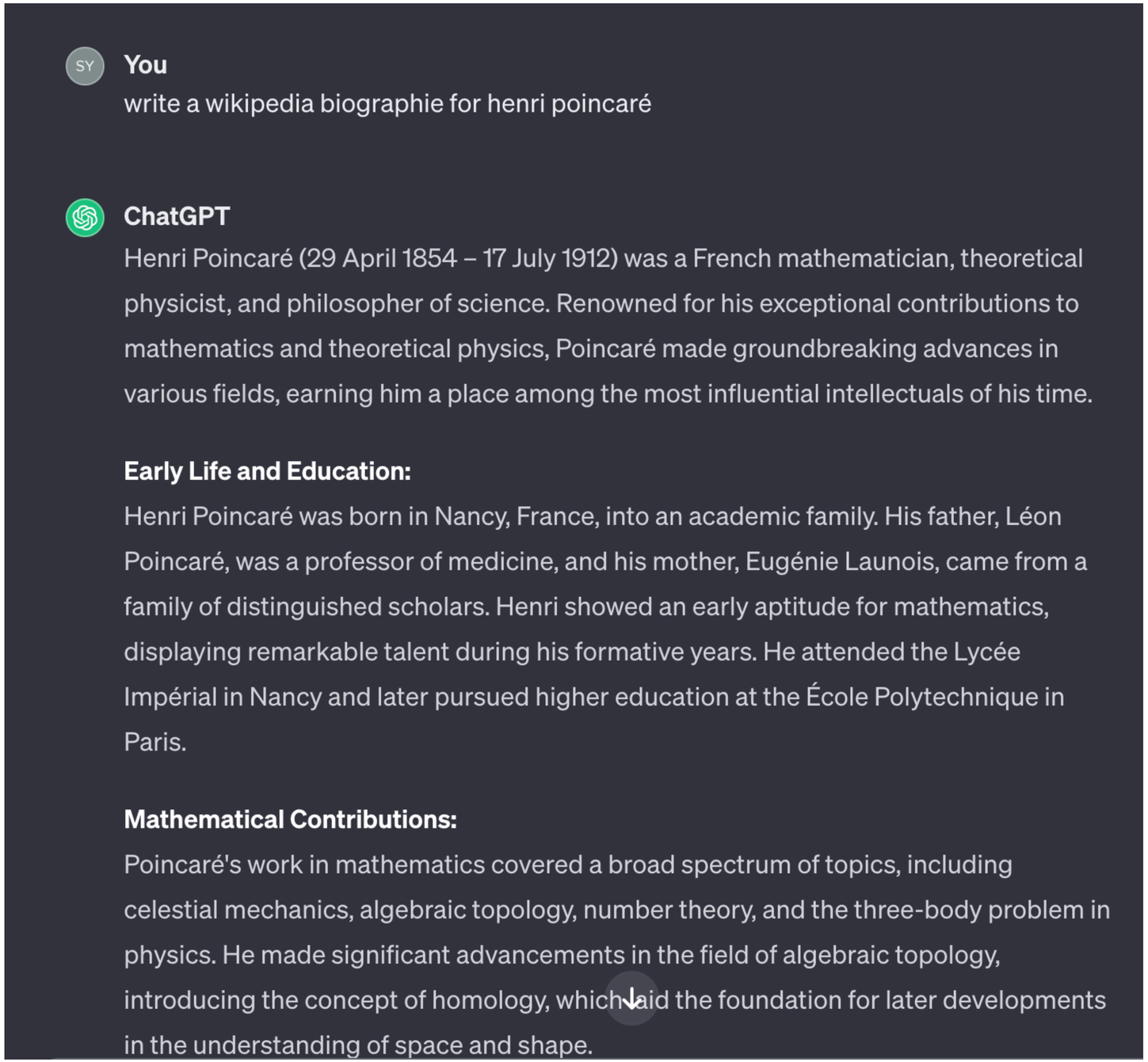
Pas Ashley McKenzie

Winner of the 2016 Toronto international film festival



ChatGPT 3.5
29/12/2023

ChatGPT connait Henri Poincaré



You
write a wikipedia biographie for henri poincaré

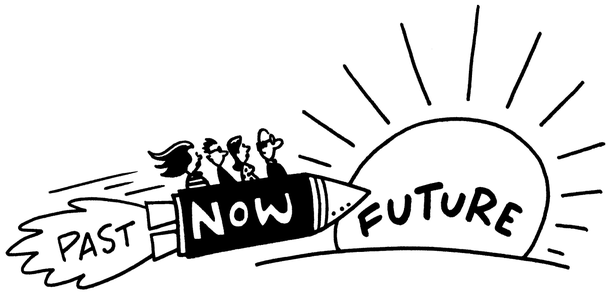
ChatGPT
Henri Poincaré (29 April 1854 – 17 July 1912) was a French mathematician, theoretical physicist, and philosopher of science. Renowned for his exceptional contributions to mathematics and theoretical physics, Poincaré made groundbreaking advances in various fields, earning him a place among the most influential intellectuals of his time.

Early Life and Education:
Henri Poincaré was born in Nancy, France, into an academic family. His father, Léon Poincaré, was a professor of medicine, and his mother, Eugénie Launois, came from a family of distinguished scholars. Henri showed an early aptitude for mathematics, displaying remarkable talent during his formative years. He attended the Lycée Impérial in Nancy and later pursued higher education at the École Polytechnique in Paris.

Mathematical Contributions:
Poincaré's work in mathematics covered a broad spectrum of topics, including celestial mechanics, algebraic topology, number theory, and the three-body problem in physics. He made significant advancements in the field of algebraic topology, introducing the concept of homology, which laid the foundation for later developments in the understanding of space and shape.

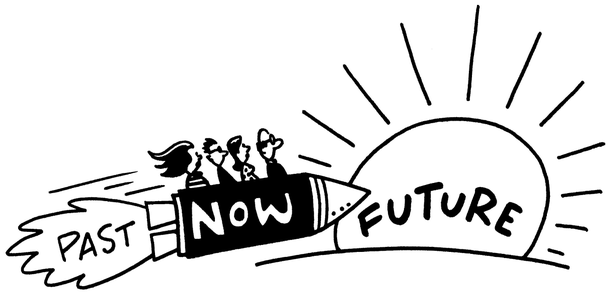
Conclusion

La génération à partir de graphes de connaissances reste limitée



- Généralisation limitée
- Multilingualité restreinte
- Métrique de similarité graphe/text multilingue pour
  - Filtrer les données bruitées
  - Guider la génération
  - Généraliser à d’autres langues et d’autres domaines

La génération vers des langues peu dotées reste un problème ouvert



- Manque de données
- Modèles inadaptés (en cascade, affinage)

Merci !